Aplicaciones de la Inferencia Bayesiana en el Análisis de Datos

Dra. Carmen Sánchez Gil

Departamento de Estadística e I.O.
Facultad de Ciencias, Universidad de Cádiz

Curso 2016/17

contents

Regla de Bayes

- Inferencia sobre parámetro θ, dado datos x: proporcionar un modelo de la *distribución de probabilidad conjunta* para θ y x

$$p(\theta, x) = p(\theta)p(x|\theta)$$

donde $p(\theta)$ es la dist. *prior*, y $p(x|\theta)$ la *dist. de la muestra (datos).*

- *D. posterior* de θ es por tanto,

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(\theta)p(x|\theta)}{p(x)}$$

- $p(x) = \sum_\theta p(\theta)p(x|\theta)$, caso discreto, o $p(x) = \int p(\theta)p(x|\theta)d\theta$ caso cont. No depende de θ, y dado x fijo $p(x) \equiv cte$
- *D. posterior no normalizada*:

$$p(\theta|x) \propto p(\theta)p(x|\theta)$$

Predicción

- Inferencia predictiva sobre el observable desconocido, antes de obtener los datos x. Dist. del obs. x es la *D. prior predictiva*:

$$p(x) = \int p(\theta, x) d\theta = \int p(\theta) p(x|\theta) d\theta$$

donde $p(\theta)$ es la dist. *prior*, y $p(x|\theta)$ la *dist. de la muestra (datos).*

- Una vez tomados los datos $x = (x_1, \ldots, x_n)$, *D. posterior predictiva* del observ. desconocido $\tilde{x}$:

$$\begin{aligned} p(\tilde{x}|x) &= \int p(\tilde{x}, \theta|x) d\theta \\ &= \int p(\tilde{x}|\theta, x) p(\theta|x) d\theta \\ &= \int p(\tilde{x}|\theta) p(\theta|x) d\theta \quad \text{(Sup. indp. de } x \text{ y } \tilde{x} \text{ dado } \theta\text{)} \end{aligned}$$

Notación

- Datos x sólo afectan d. post. $p(\theta|x)$ a través de $p(x|\theta)$. Para x fijo, $\mathcal{L}(\theta) = p(x|\theta)$ *f. verosimilitud*.
- Dados u, v: $p(u, v)$ *densidad conjunta*, $p(u|v)$ *dist. condicional*, y $p(u) = \int p(u, v)dv$ *densidad margial*.
- *Factorización d. conj.* como producto de d. marginal y condicional: $p(u, v, w) = p(u|v, w)p(v|w)p(w)$
- Media y var. de dist. cond.:

$$\begin{aligned} E(u) &= E(E(u|v)) \\ Var(u) &= E(Var(u|v)) + Var(E(u|v)) \end{aligned}$$

- Transformación de variables: $v = f(u)$, f función $1-1$

$$\begin{aligned} p_v(v) &= p_u(f^{-1}(v)) \quad \text{caso discreto} \\ p_v(v) &= |J|p_u(f^{-1}(v)) \quad \text{caso continuo} \end{aligned}$$

Ejm: $u \in (0, \infty)$, $log(u) \in \mathcal{R}$; $u \in [0, 1]$, $logit(u) = log\left(\frac{u}{1-u}\right) \in \mathcal{R}$

Priors no informativas

- No base poblacional. Mínima función/papel sobre la d. post.
- Principio de Invarianza de Jeffreys: aprox. para def. d. prior no inform., basadas en transfromaciones $1-1$ del param. $\phi = h(\theta)$.

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| = p(\theta)\,|h'(\theta)|^{-1}$$

- Aplicando PIJ, D. prior no informativa: $\boxed{p(\theta) \propto J(\theta)^{1/2}}$, donde $J(\theta) = -E[\frac{d^2 logp(x|\theta)}{d\theta^2}|_\theta]$ Información de Fisher de θ.
- Modelo de prior de Jeffreys es invariante respecto parametrizaciones:

$$\begin{aligned} J(\phi)^{1/2} &= \left(-E\left[\frac{d^2 logp(x|\phi)}{d\phi^2}\right]\right)^{\frac{1}{2}} \\ &= \left(-E\left[\frac{d^2 logp(x|\theta = h^{-1}(\phi))}{d\theta^2}\left|\frac{d\theta}{d\phi}\right|^2\right]\right)^{\frac{1}{2}} \\ &= \left(J(\theta)\left|\frac{d\theta}{d\phi}\right|^2\right)^{\frac{1}{2}} = J(\theta)^{1/2}\left|\frac{d\theta}{d\phi}\right| \end{aligned}$$

Priors no informativas: Pivotes

Aprox. donde se toma la distr. del pivote como su d. post. Se puede aplicar a estad. suficientes (modelos jerárquicos)

(1) θ param. localización puro, y $u = x - \theta$ pivote si densidad x: $p(x - \theta|\theta) = f(u)$, i.e. libre de θ y x

- Si $p(\theta)$ prior no informativa, $f(u)$ lo será para la d. post $p(u|x)$ $\rightarrow x - \theta$ pivote de d. post.
- Como (R. Bayes) $p(x - \theta|x) \propto p(\theta)p(x - \theta|\theta)$, necesariamente densidad priori no informativa es uniforme en θ: $p(\theta) \propto cte$ para $\theta \in (-\infty, \infty)$

Priors no informativas: Pivotes

(2) θ param. escala puro, y $u = \frac{x}{\theta}$ pivote si densidad x: $p(\frac{x}{\theta}|\theta) = g(u)$, i.e. libre de θ y x

- Si $p(\theta)$ prior no informativa, $g(u)$ lo será para la d. post $p(u|x)$
- Aplicando transf. var.,

$$\left.\begin{array}{rcccl} p(x|\theta) & = & \frac{1}{\theta}p(u|\theta) & = & \frac{1}{\theta}g(u) \\ p(\theta|x) & = & \frac{x}{\theta^2}p(u|x) & = & \frac{x}{\theta^2}g(u) \end{array}\right\} \Rightarrow p(\theta|x) = \frac{x}{\theta}p(x|\theta)$$

de donde, $p(\theta) \propto \frac{1}{\theta}$

- Equiv. ($\phi = h(\theta)$): $p(log\theta) \propto 1$, o $p(\theta^2) \propto \frac{1}{\theta^2}$

Estimar una proporción

- Población de estudio: estudiantes universitarios
- Los estudiantes no duermen suficientes horas.
- $\theta =$ proporción de estudiantes que duermen al menos 8 h
- Un estudo dice que $\theta < 0.5$. ¿Valor real de θ ?
- Una muestra de 27 estudiantes, 11 dormían al menos 8h
- Si tomamos nueva muestra de 20 estudiantes, ¿θ?

- Datos: muestra de tamaño $n = 27$, $s = 11$ estudiantes duermen al menos 8h (éxitos). Por tanto, $f = 16$ no lo hacen (fracasos), y $\bar{\theta} = 11/27$ proporción muestral.
- Densidad de la muestra: $\theta \sim Beta(\alpha, \beta)$, $\theta \in [0, 1]$

$$p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} = \mathcal{L}(\theta) \quad \text{(Verosimilitud)}$$

$\alpha - 1 \equiv s$, nº éxitos, y $\beta - 1 \equiv f$, nº fracasos $\Rightarrow$ Beta(12,17)

- Aplicando la regla de Bayes, la densidad posterior de θ es

$$p(\theta|datos) \propto p(\theta)p(\theta|\alpha, \beta)$$

donde $p(\theta)$ d. prior de θ

Prior discreta

- Lista de posibles valores de la proporción: $\{\theta_1, \ldots \theta_n\}$

$$0.05, 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95$$

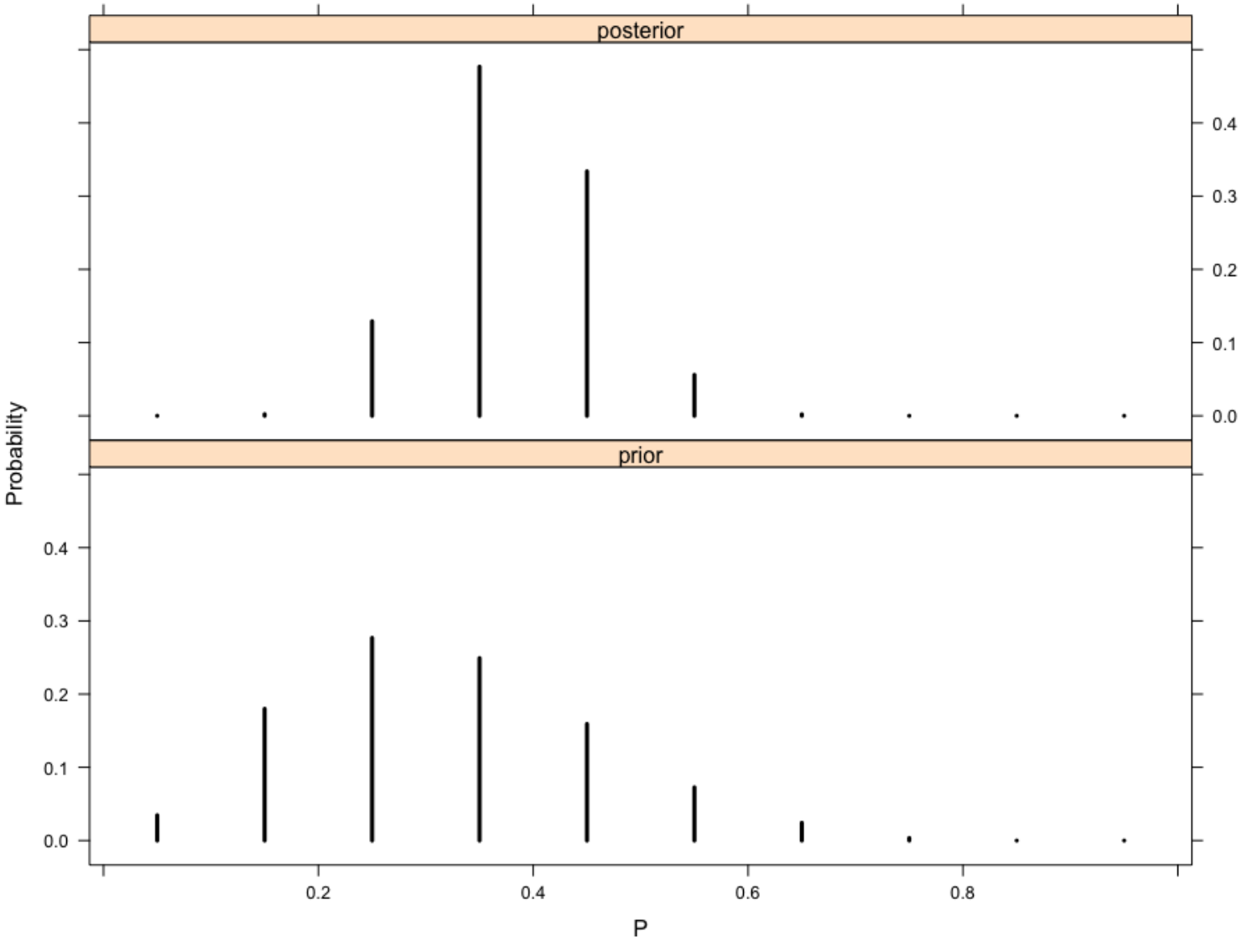
- Y sus pesos correspondientes: $\{w_1, \ldots w_n\}$

$$1, 5.2, 8, 7.2, 4.6, 2.1, 0.7, 0.1, 0, 0$$

- D. post de θ

$$p(\theta|datos) \propto \sum_{i=1}^{n} p_i \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}$$

donde $p_i = \frac{w_i}{\sum_{i=1}^{n} w_i}$, $\alpha = s + 1 = 12$ y $\beta = f + 1 = 17$


posterior
prior
Probability
P
0.0
0.1
0.2
0.3
0.4
0.2
0.4
0.6
0.8

Prior Beta

- Como θ es continua, construimos d. prior en (0,1) que contenga las hipótesis o conocimientos a priori.
- Además de la misma forma funcional que la verosimilitud: $Beta(a, b)$

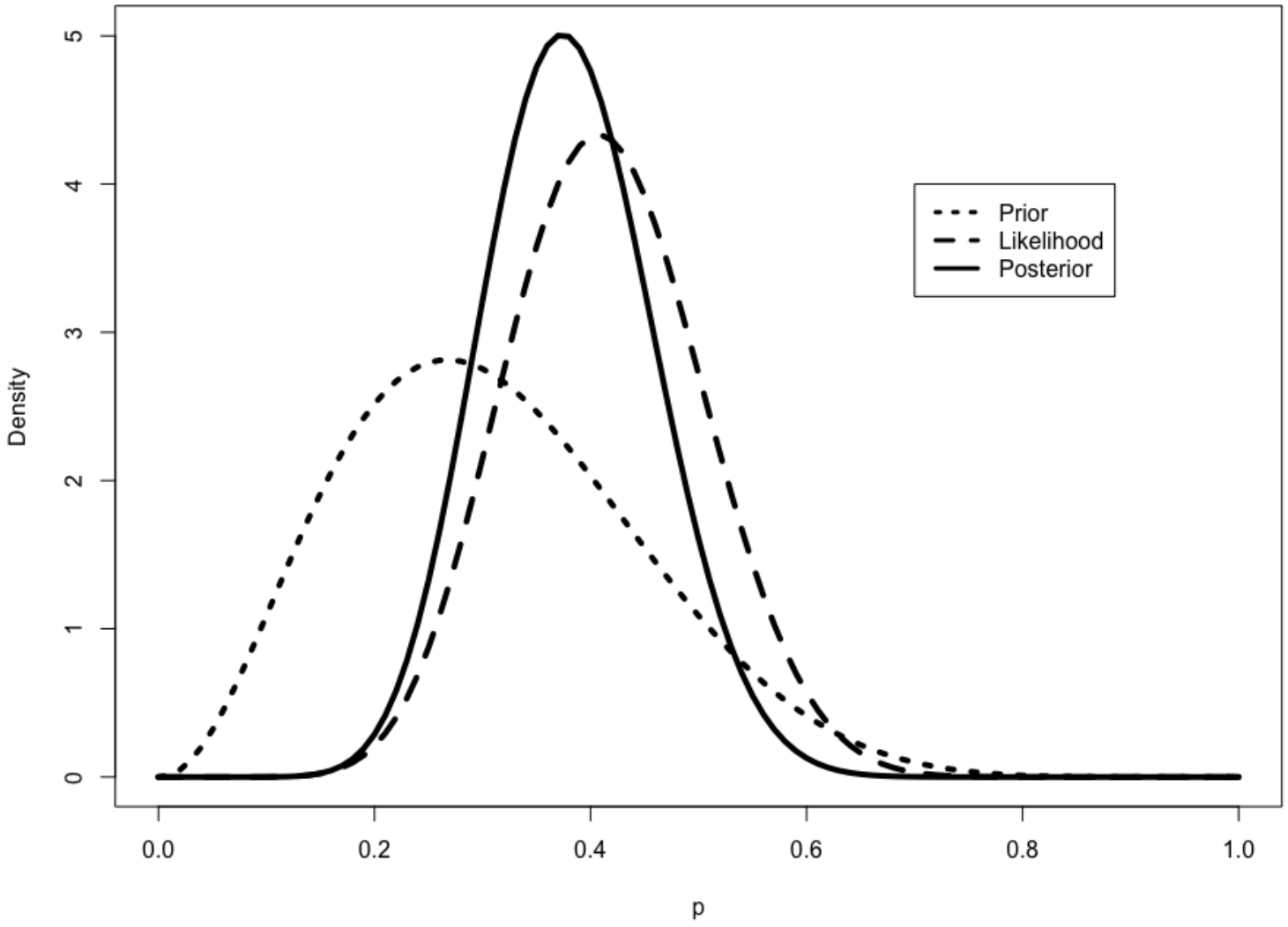
$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}, \quad 0 < \theta < 1$$

donde los hiperparámetros (a, b) contengan/reflejen información previa. $E(\theta) = \frac{a}{a+b}$, $Var(\theta) = \frac{ab}{(a+b)^2(a+b+1)}$

- De forma indirecta a través percentiles de la distrib: $P(\theta < 0.3) = 0.50$ y $P(\theta < 0.5) = 0.90 \Rightarrow (a = 3.26, b = 7.19)$
- D. post de θ sigue también una $Beta(\alpha + a - 1, \beta + b - 1)$

$$p(\theta|datos) \propto \theta^{\alpha+a-2}(1-\theta)^{\beta+b-2}$$

- Conjugadas: prior y post. tienen misma forma funcional.


Density
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
p
Prior
Likelihood
Posterior

- El uso de una prior $Beta(a, b)$ tiene ventajas computacionales, pues obtenemos la forma analítica de la d. post.
- Otra alternativa: simulación computacional de la d. post, dada una prior $p(\theta)$ arbitraria.
  1. Determinar un *grid* de valores del parámetro/s $\{\theta_i\}_{i=1,\ldots,n}$, que cubran la d. post.
  2. Aproximación de la d. post por una distr. prob. discreta:
     - Calculamos el producto de la prior, $p(\theta)$, por la verosimilitud, $\mathcal{L}(\theta)$, sobre el grid: $p(\theta_i|datos) \propto p(\theta_i)\mathcal{L}(\theta_i)$, $\forall i$.
     - Normalizamos (covertimos estos prod. en prob.):

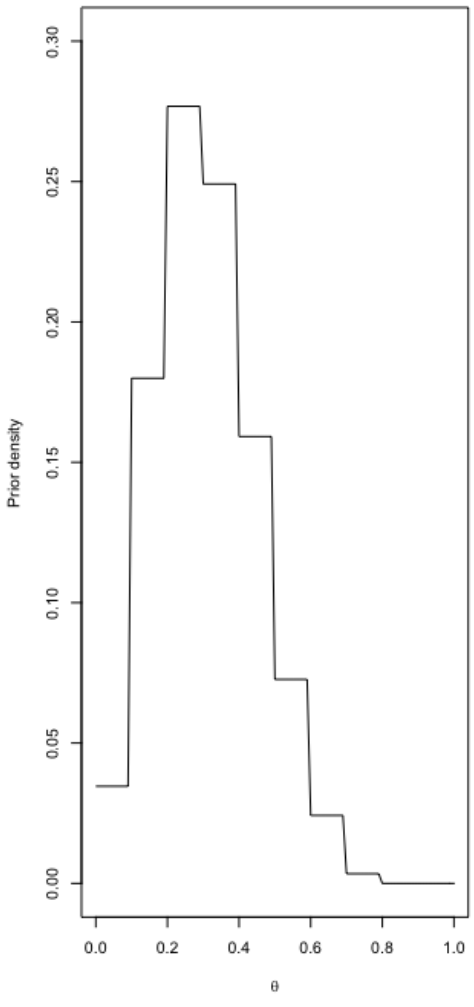
$$p(\theta_i|datos) \sim \frac{p(\theta_i)\mathcal{L}(\theta_i)}{\sum_{j=1}^{n} p(\theta_j)\mathcal{L}(\theta_j)}$$

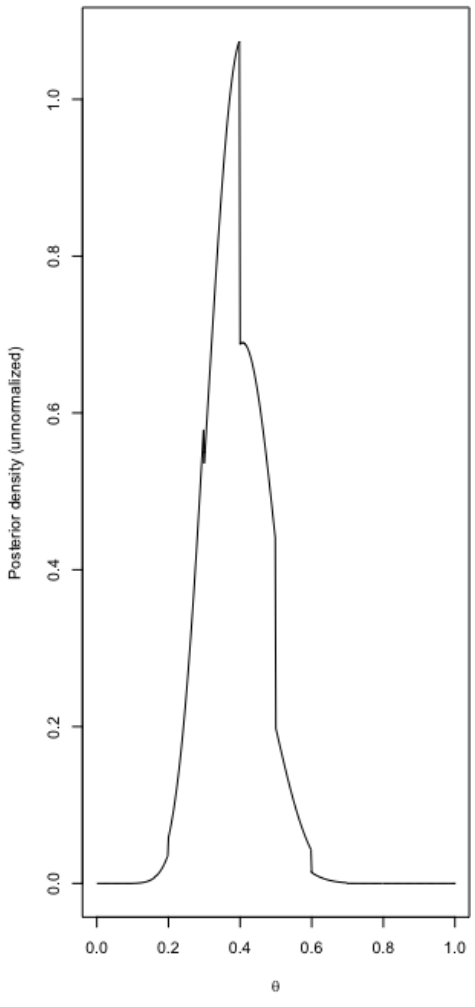
  3. Tomamos una m.a.s. de dicha distrb. discreta, que resuma la d. post. del param.: $\{\theta_1, \ldots, \theta_m \mid data\}$

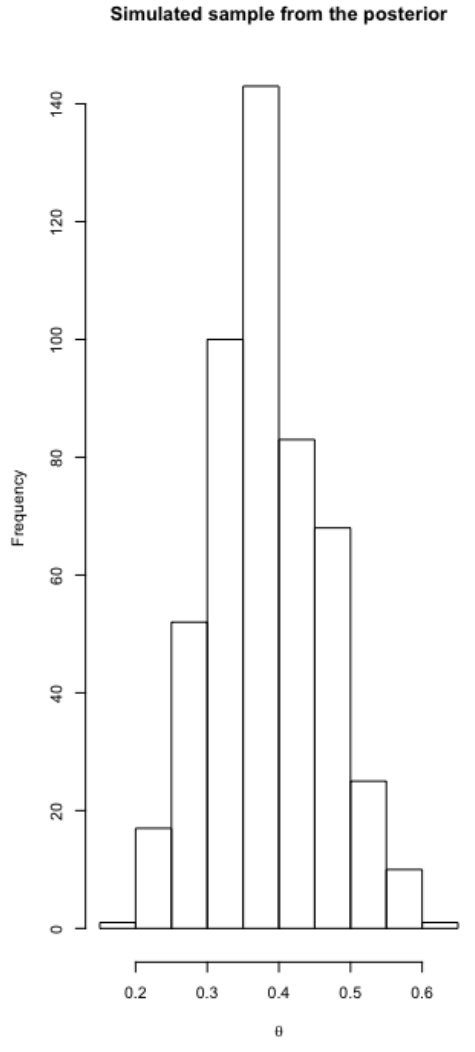
Prior "Histograma"

- Supongamos que expresamos nuestro conocimiento a priori sobre θ como un histograma.
- Grid igualmente espaciado: dividimos el $rg(\theta) = [0, 1]$, en 10 subintervalos equidistantes: $(0, 0.1), \ldots, (0.9, 1)$.
  $\Rightarrow \{\theta_i\} = \{0.05, 0.15, \ldots, 0.95\}$ ptos. medios de sub-interv.
- La d. prior se define entonces como $p_i = w_i / \sum_{j=1}^{n} w_j$, dados los pesos $\{w_1, \ldots, w_{10}\} = \{1, 5.2, 8, 7.2, 4.6, 2.1, 0.7, 0.1, 0, 0\}$.
- Recordamos que la f. verosimilitud para la proporción de una binomial es $\mathcal{L}(\theta) = Beta(\alpha, \beta)$, donde $\alpha = s + 1$ y $\beta = f + 1$.
- D. post viene dada entonces por:

$$p(\theta_i|datos) \sim \frac{w_i}{\sum_{j=1}^{n} w_j} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1 - \theta_i)^{\beta-1}$$


Prior density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.2
0.4
0.6
0.8
1.0
θ


Posterior density (unnormalized)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
θ


Simulated sample from the posterior
Frequency
0
20
40
60
80
100
120
140
0.2
0.3
0.4
0.5
0.6
θ

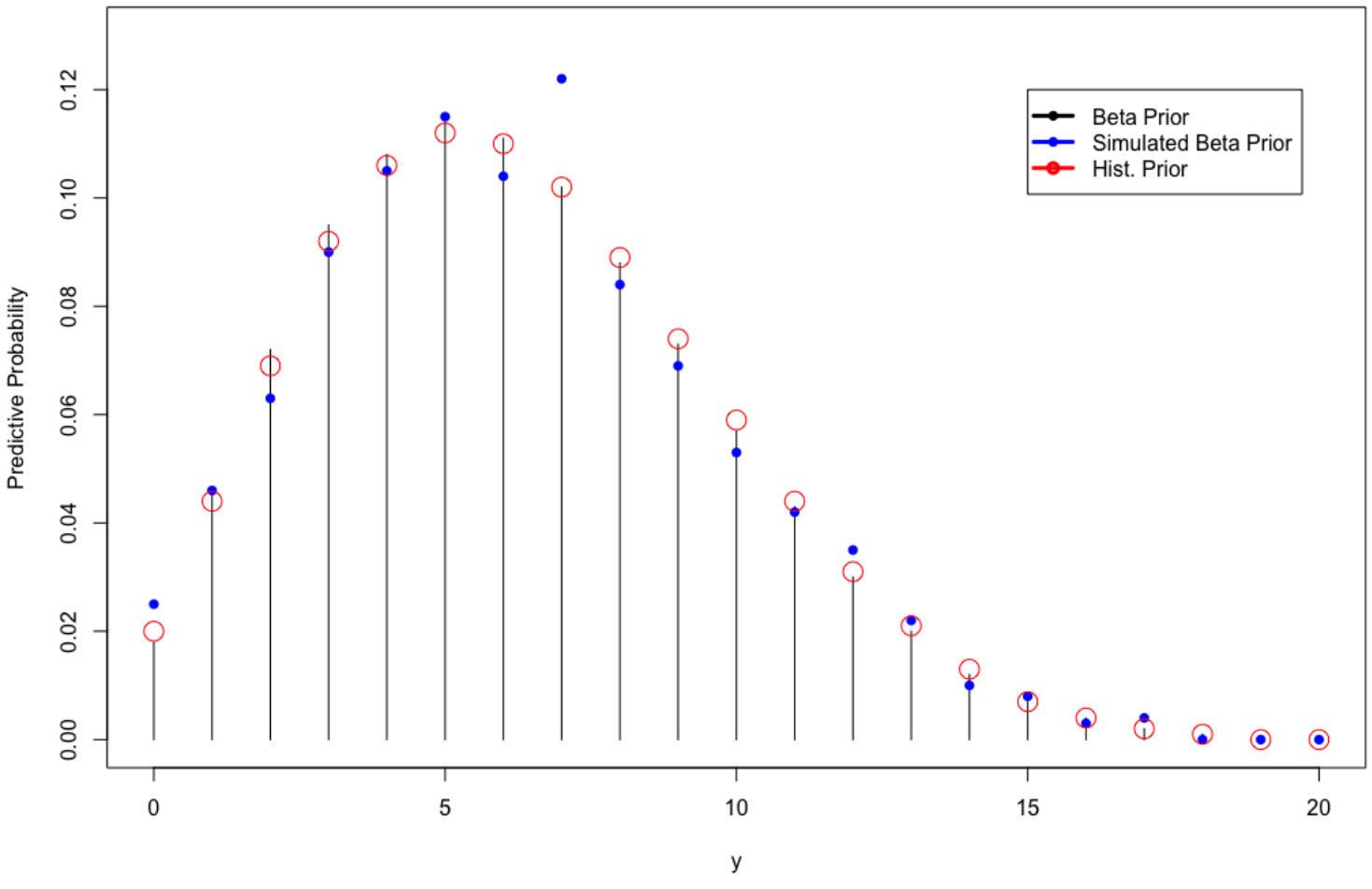
D. Predictiva

- Queremos predecir el número de estudiantes que duermen más de 8h, $\tilde{x}$, en una futura muestra de $n = 20$ estudiantes.
- D. prior predictiva: $p(x) = \int p(\theta, x)d\theta = \int p(\theta)p(x|\theta)d\theta$
- Verosimilitud: $\mathcal{L}(\theta) = p_{Bin}(x \mid n, \theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, $x = 0, \ldots, n$
- Prior discreta: $\{\theta_1, \ldots \theta_m\}$, con prob. $\{p_1, \ldots p_m\}$

$$p(\tilde{x}) = \sum_{i=1}^{m} p_i \binom{n}{\tilde{x}} \theta_i^{\tilde{x}}(1-\theta_i)^{n-\tilde{x}}$$

- Prior Beta(a,b):

$$\begin{aligned} p(\tilde{x}) &= \int_0^1 \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}\binom{n}{\tilde{x}}\theta^{\tilde{x}}(1-\theta)^{n-\tilde{x}}d\theta \\ &= \binom{n}{\tilde{x}}\frac{B(a+\tilde{x}, b+n-\tilde{x})}{B(a,b)} \quad \tilde{x} = 0, \ldots, n \end{aligned}$$


Beta Prior
Simulated Beta Prior
Hist. Prior
Predictive Probability
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
5
10
15
20
y

Resumen de las inferencias por Simulación computacional

Parte central de las aplicaciones del análisis Bayesiano.
Podemos muestrear distr. prob, incluso cuando la f. densidad no pueda ser integrada explícitamente.

1. Determinar un *grid* de valores del parámetro/s $\{\theta_i\}_{i=1,\ldots,n}$, que cubran la d. post.
2. Aproximación de la d. post por una distr. prob. discreta:
   - Calculamos el producto de la prior, $p(\theta)$, por la verosimilitud, $\mathcal{L}(\theta)$, sobre el grid: $p(\theta_i|datos) \propto p(\theta_i)\mathcal{L}(\theta_i)$, $\forall i$.
   - Normalizamos (covertimos estos prod. en prob.):

$$p(\theta_i|datos) \sim \frac{p(\theta_i)\mathcal{L}(\theta_i)}{\sum_{j=1}^{n} p(\theta_j)\mathcal{L}(\theta_j)}$$

3. Tomamos una m.a.s. de dicha distrb. discreta, que resuma la d. post. del param.: $\{\theta_1, \ldots, \theta_m \mid data\}$ (m=1000)

Muestreo a través de la inversa de la F.d.D

- Dada una distr. $p(x)$, $F(a) = P(x \leq a) = \begin{cases} \sum_{x \leq a} p(x) & \text{discreta} \\ \int_{-\infty}^{a} p(x)dx & \text{continua} \end{cases}$
- Dada $u \sim U(0,1)$, $x = F^{-1}(u)$ única, y fácil de calcular/tabular (F no necesariamente $1-1$, discreta)

Ejemplo

Caso continuo: $x \sim Exp(\lambda)$, $F(x) = 1 - e^{-\lambda x} = u \rightarrow x = F^{-1}(u) = -\frac{log(1-u)}{\lambda}$.
Si $u \sim U(0,1)$, también $(1-u) \sim U(0,1)$.
Muestreando $(u_1, \ldots, u_n) \sim U(0,1)$, obtenemos m.a.s. de d. expon. como:

$$-\frac{log(u_1)}{\lambda}, \ldots, -\frac{log(u_n)}{\lambda} \sim Exp(\lambda)$$

Caso discreto: $x_1 \leq x_2 \leq \ldots \leq x_k$, con f.masa prob $p(x_i)$. $F(x_j) = \sum_{i \leq j} p(x_i)$, una partición de $(0,1)$: $(0, F(x_1)), (F(x_1), F(x_2)), \ldots, (F(x_{k-1}), 1)$
Muestreando $u \sim U(0,1)$, entonces:

$$P(F(x_{j-1}) \leq u \leq F(x_j)) = F(x_j) - F(x_{j-1} = p(x_j) = P(x = x_j)$$

Modelos uniparamétricos: Modelo Binomial

- Proceso de *Bernouilli*.
- Datos: $x_1, x_2, \ldots, x_n$ variables binarias (0/1, éxito / fracaso)
- $x =$ num. total de éxitos en n intentos.
- θ proporción de éxito, o prob. de éxito en cada prueba.

$$p(x|\theta) = B(x|n,\theta) = \binom{n}{x}\theta^n(1-\theta)^{n-x}$$

Ejemplo: Estimar la proporción de nacimientos de mujeres, en casos de placenta previa

- Actualmente se establece que en Europa $\theta = 0.485$
- Se observa una muestra de 980 nacimientos con casos de placenta previa, donde 437 fueron niñas y 543 niños.
- Es la proporción θ la misma ??

Alternativamente podemos utilizar la razón de nacimiento de hombres respecto al de mujeres, $\phi = \dfrac{1-\theta}{\theta}$

Sea $x =$ número de niñas en n nacimientos: $x \sim B(n, \theta)$

- Para aplicar inferencia Bayesiana en el modelo binomial, necesitamos especificar la distribución prior de θ.
- El caso más sencillo, que aplicaremos de momento, es asumir una Uniforme en el intervalo $[0, 1]$: $\theta \sim U(0, 1)$.
- La distribución posterior para θ (no normalizada) resulta entonces:

$$p(\theta|x) \propto p(\theta)p(x|\theta) \propto \theta^x(1-\theta)^{n-x}$$

- El factor $\binom{n}{x}$, que no depende de θ, se puede tratar como constante en el cálculo de su distribución posterior.
- Podemos observar que dicha distrib. post. tiene la forma de una distribución *beta* sin normalizar: $\theta|x \sim Beta(x+1, n-x+1)$.

D. Post. incorpora información de los datos y la prior

- La esperanza priori de θ is el promedio de todas las posibles medias post. sobre la distr. de los datos.

$$E(\theta) = E(E(\theta|x))$$

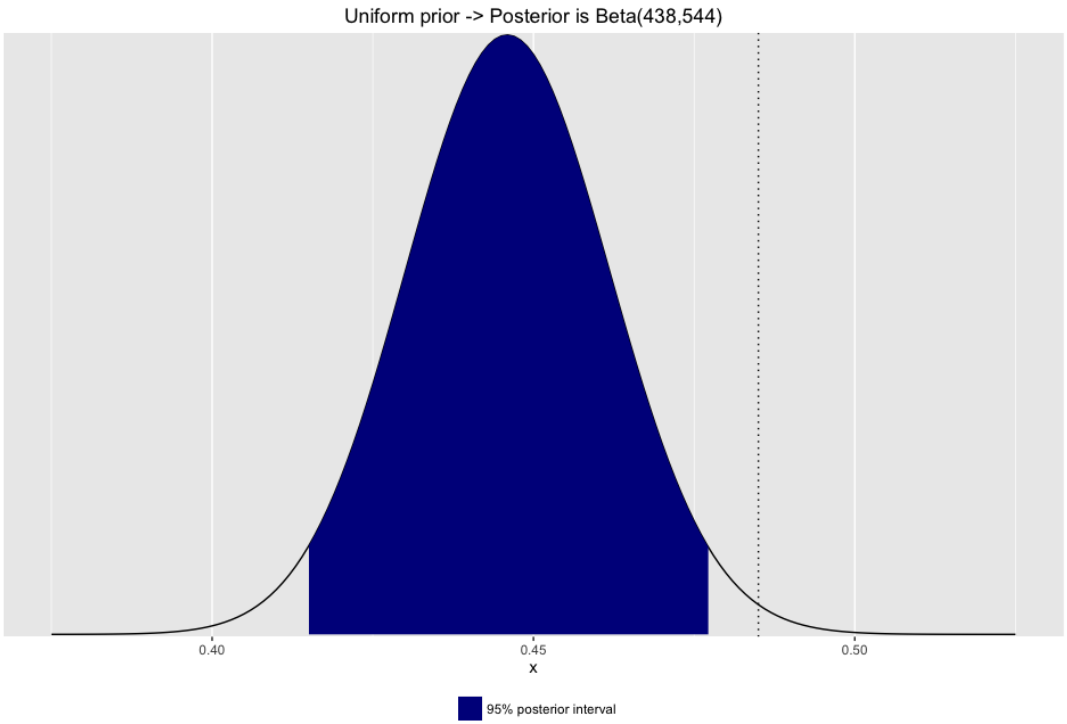
- Var. post. es en promedio menor que la var. priori, en función de la var. de la media post. sobre distr. de los datos.

$$Var(\theta) = E(Var(\theta|x)) + Var(E(\theta|x))$$

- $E(\theta) = \frac{a+b}{2} = \frac{1}{2}$, $Var(\theta) = \frac{(b-a)^2}{12} = \frac{1}{12}$ (Uniforme)
- $E(\theta|x) = \frac{\alpha}{\alpha+\beta} = \frac{x+1}{n+2}$ (Beta), donde $E(\theta) = \frac{1}{2}$ y $\bar{\theta} = \frac{x}{n}$.

Ejemplo: Estimar la proporción de nacimientos de mujeres, en casos de placenta previa

- $\theta = 0.485$
- $\bar{\theta} = 437/980$ niñas, y $1 - \bar{\theta} = 543/980$ niños ($n = 980$).



Familia de Distribuciones a priori (conjugadas)

- Si expresamos la verosimilitud en términos generales como $p(x|\theta) \propto \theta^a(1-\theta)^b$, una familia de distr. prior con la misma forma

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \theta \sim Beta(\alpha, \beta)$$

  $\alpha - 1$ aciertos a priori, y $\beta - 1$ fallos $\rightarrow$ *hiperparámetros*

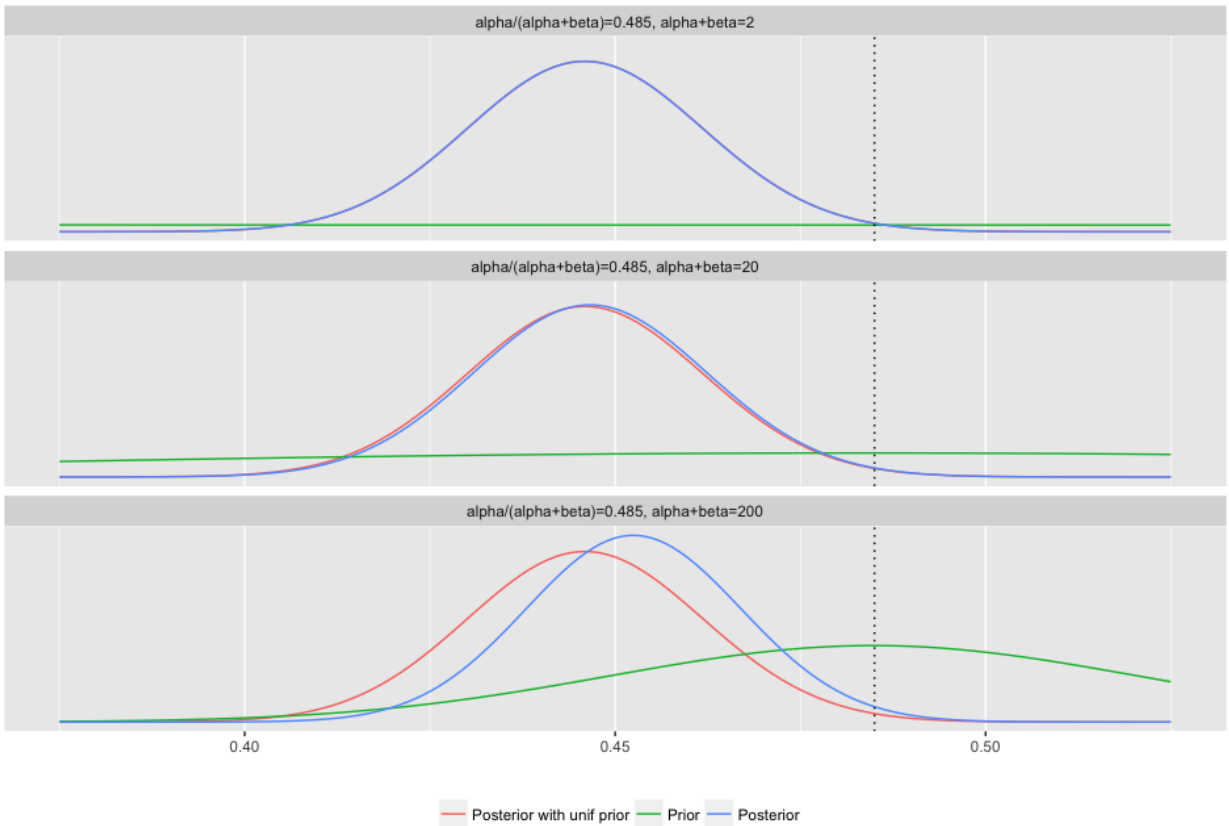
- La distrib. post. de θ tiene la misma forma que la prior: *conjugadas*

$$\begin{aligned} p(\theta|x) &\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \\ &= Beta(\theta|\alpha+x, \beta+n-x) \end{aligned}$$

- $Beta(\alpha, \beta)$ es una *familia conjugada* de la $B(n, \theta)$.

Ejemplo: Estimar la proporción de nacimientos de mujeres, en casos de placenta previa

- $\theta = 0.485$; $\bar{\theta} = 437/980$ niñas.
- $\alpha + \beta - 2 \sim$ num. observ. priori



- La Uniforme es un caso particular: $U(0,1) \equiv Beta(1,1)$
- $E(\theta|x) = \dfrac{\alpha + x}{\alpha + \beta + n}$, toma valores entre la proporción muestral $\frac{x}{n}$ y la media priori, $\frac{\alpha}{\alpha+\beta}$
- $Var(\theta|x)\dfrac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} = \dfrac{E(\theta|x)(1 - E(\theta|x))}{\alpha + \beta + n + 1}$
- Para α, β fijos, $\uparrow x$, $\uparrow (n - x)$: $E(\theta|x) \approx \frac{x}{n}$ y $Var(\theta|x) \approx \frac{1}{n}\frac{x}{n}(1 - \frac{x}{n})$
- En el límite, los parámetros de la distr. prior no tienen influencia sobre la distr. post.
- TCL: $\left(\dfrac{\theta - E(\theta|x)}{DT(\theta|x)}\middle| x\right) \to N(0,1)$

Media Normal, con varianza conocida

- $x \sim N(\mu, \sigma)$: $p(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
- Familia de priors conjugada: $\mu \sim N(\mu_0, \tau_0)$, $p(\mu) \propto e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2}$
- D. post. de μ

$$\begin{aligned} p(\mu|x) &\propto p(\mu)p(x|\mu) \propto exp\left(-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right) exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\ &\propto exp\left(-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\mu-\mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\right) \\ &\propto exp\left(-\frac{1}{2\tau_1^2}(\mu-\mu_1)^2\right) \end{aligned}$$

$$p(\mu|x) = N(\mu|\mu_1, \tau_1), \text{ donde } \mu_1 = \frac{\frac{\mu_0}{\tau_0^2} + \frac{1}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}, \text{ y } \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

Media Normal, con varianza conocida

- D. post. pred. $\tilde{x}$

$$\begin{aligned} p(\tilde{x}|x) &= \int p(\tilde{x}|\theta)p(\theta|x)d\theta \\ &\propto \int exp\left(-\frac{1}{2\sigma^2}(\tilde{x}-\theta)^2\right) exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right) d\theta \end{aligned}$$

  - $p(\tilde{x}|x)$ no depende de x;
  - D. post. conj. de $(\tilde{x}|x)$ es normal $\rightarrow$ d. post. marg. $\tilde{x}|\theta$ normal
  - $E(\tilde{x}|x) = E(E(\tilde{x}|\theta, x)|x) = E(\theta|x) = \mu_1$
  - $Var(\tilde{x}|x) = E(Var(\tilde{x}|\theta, x)|x) + Var(E(\tilde{x}|\theta, x)|x) =$
    $= E(\sigma^2|x) + Var(\theta|x) = \sigma^2 + \tau_1^2$

Media Normal, con varianza conocida

- $x = (x_1, x_2, \ldots, x_n)$ i.i.d., $x_i \sim N(\mu, \sigma)$: $p(x_i|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$
- Familia de priors conjugada: $\mu \sim N(\mu_0, \tau_0)$, $p(\mu) \propto e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2}$
- D. post. de μ

$$\begin{aligned} p(\mu|x) &\propto p(\mu)p(x|\mu) = p(\mu)\prod_{i=1}^{n} p(x_i|\mu) \\ &\propto exp\left(-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2\right)\prod_{i=1}^{n} exp\left(-\frac{1}{2\sigma^2}(x_i-\mu)^2\right) \\ &\propto exp\left(-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\mu-\mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)\right) \end{aligned}$$

$$p(\mu|x_1, \ldots, x_n) = p(\mu|\bar{x}) = N(\mu|\mu_n, \tau_n);\ \mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}},\ \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2},$$

y $\bar{x} = t(x)$ *estadístico suficiente*

Media Normal, con varianza conocida

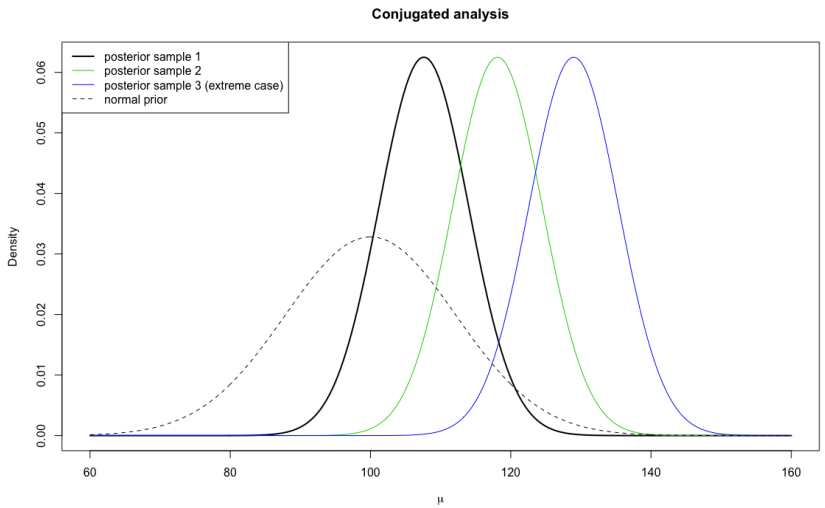
- $\mu_n = \bar{x} - (n\bar{x} - \mu_0)\dfrac{\sigma^2}{\sigma^2 + n\tau_0^2}$
- Casos extremos, d. post. dominada por $\bar{x}$ y σ^2: $p(\mu|x) \sim N(\bar{x}, \frac{\sigma^2}{n})$
  - $n \gg$ grande
  - $\tau_0^2 = \sigma^2$
  - τ_0 fija, $n \to \infty$; o n fija, $\tau_0 \to \infty$
  - $1/\tau_0^1 \ll n/\sigma^2$
- D. post. es aproximadamente como si $p(\mu) \propto cte$, para $\mu \in (-\infty, \infty)$!! $\to$ *Prior Impropia*

> *$p(\theta)$ prior propia si no depende de los datos y su integral vale 1 (cte, renormalizamos)*

- Prior es impropia, pero D. post. es propia

Estimar μ Normal, con σ conocida

- Fichero *achievement* del paquete `LearnBayes`: datos 109 observaciones para un grupo niños de un colegio Austríaco.
- $X = IQ \sim N(\mu, \sigma)$, asumimos $\sigma = 15 \Rightarrow \bar{X} \sim N(\mu, 15/\sqrt{n})$
- Conocimiento priori: $P_{50} = 100$ y $P_{95} = 120$
- Prior conjugada $\mu \sim N(\mu_0, \tau_0^2)$: $\mu \mid x_1, \ldots, x_n \sim N\left(\dfrac{\frac{\mu_0}{\tau_0^2} + \frac{n}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \dfrac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}\right)$

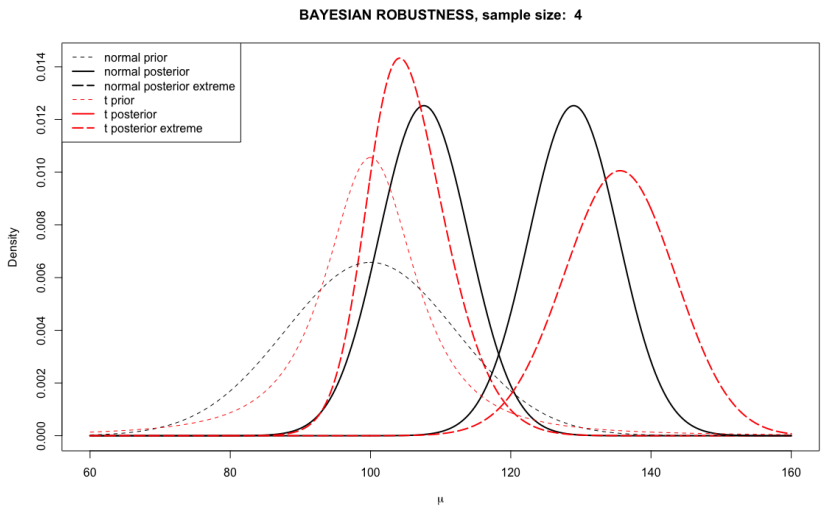


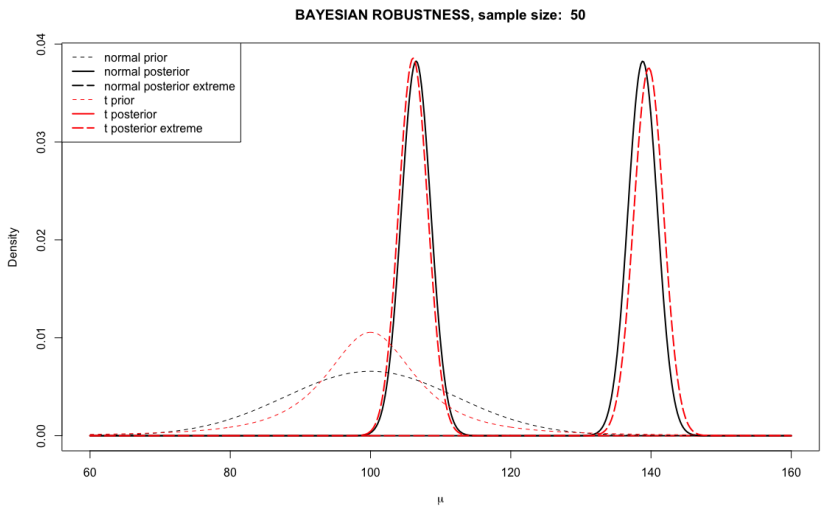
Estimar μ Normal, con σ conocida: Robustez bayesiana

- Prior no conjugada $\mu \sim t_{\nu_0}(\mu_0, \tau_0)$: $p(\mu) \propto \left(1 + \frac{1}{\nu_0}\left(\frac{\mu-\mu_0}{\tau_0}\right)^2\right)^{-(\nu_0+1)/2}$

$$\begin{aligned} p(\mu|x) &\propto p(\mu)p(\bar{x}|\mu) \\ &\propto \left(1 + \frac{1}{\nu_0}\left(\frac{\mu-\mu_0}{\tau_0}\right)^2\right)^{-(\nu_0+1)/2} exp\left(-\frac{n}{2\sigma^2}(\bar{x}-\mu)^2\right) \end{aligned}$$

No tiene forma funcional "conveniente" (conocida): Aproximamos d. cont. por d. discreta sobre grid $\{\mu_1, \ldots, \mu_k\}$: $p(\mu_j|x) \propto p(\mu_j)p(\bar{x}|\mu_j)$





Varianza Normal, con media conocida

- $x = (x_1, \ldots, x_n)$ i.i.d., $x_i \sim N(\mu, \sigma)$: $p(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$
- Verosimilitud (μ conocida)

$$p(x|\sigma) \propto \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) = (\sigma^2)^{-n/2} e^{-\frac{n}{2\sigma^2}\nu}$$

$\nu = S^2$ *estadístico suficiente* ($p(\sigma|\nu,x) = p(\sigma|\nu)$)

- Prior conjugada: *Inv*-$\Gamma(\alpha,\beta)$, $p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2}$, o equiv. $\sigma^2 \sim$ Scaled Inv-$\chi^2(\nu_0, \sigma_0^2)$, $(\alpha = \nu/2, \quad \beta = \nu\sigma^2/2)$
- Distr. post.

$$\begin{aligned} p(\sigma^2|x) &\propto p(\sigma^2)\cdot p(x|\sigma^2) \propto \left(\frac{\sigma_0^2}{\sigma^2}\right)^{\frac{\nu_0}{2}+1} e^{-\frac{\nu_0\sigma_0^2}{2\sigma^2}} \cdot (\sigma^2)^{-\frac{n}{2}} e^{-\frac{n}{2\sigma^2}\nu} \\ &\propto (\sigma^2)^{-\left(\frac{n+\nu_0}{2}+1\right)} exp\left(-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2 + n\nu\right)\right) \end{aligned}$$

Varianza Normal, con media conocida

$$\sigma^2|x \sim \textit{Scaled Inv} - \chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + n\nu}{\nu_0 + n}\right)$$

- Si ν_0 (gl prior) $\ll$ n (gl datos) $\rightarrow$ $p(\sigma^2|x) \sim \textit{Scaled Inv} - \chi^2(n, \nu)$
- Esta forma de D. post también resulta definiendo una prior para σ^2 como $p(\sigma^2) \propto 1/\sigma^2$, impropia (!)
- Priors impropias pueden proporcionar d. post. propias
- Sin embargo, d. post. obtenidas de priors impropias deben interpretarse con sumo cuidado. Son aproximaciones donde $\mathcal{L}$ domina f.d.d. prior.
- Priors impropias, ejemplos de priors no informativas

Estimar σ Normal, con μ conocida

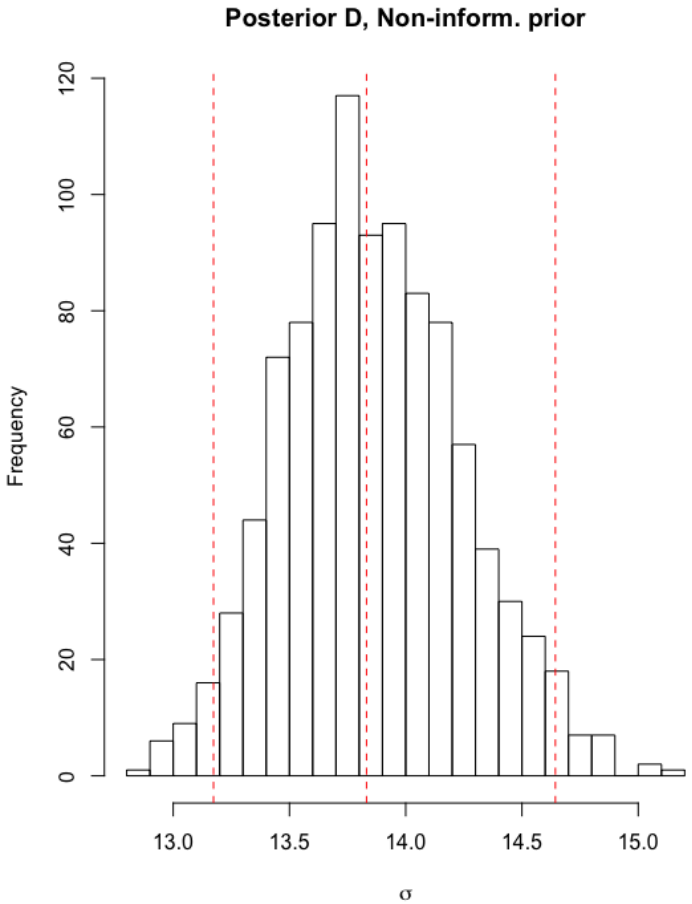
- Puntuaciones fútbol americano, X = diferencia entre resultado de un partido (winning score − losing score) frente a la diferencia de puntos publicados.
- $X \sim N(0, \sigma)$
- Fichero *footballscores* del paquete `LearnBayes`
- Prior conjugada $\sigma^2 \sim$ Scaled Inv-$\chi^2(\nu_0, \sigma_0^2)$:

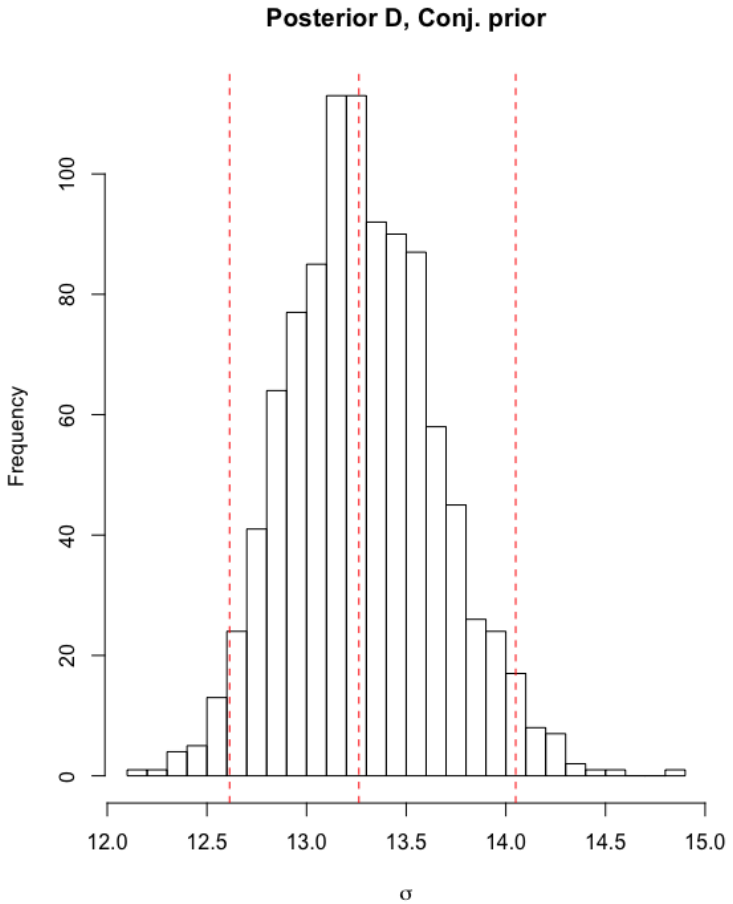
$$\sigma^2|x \sim Scaled\ Inv - \chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + n\nu}{\nu_0 + n}\right)$$

- Prior no informativa $p(\sigma^2) \propto 1/\sigma^2$:

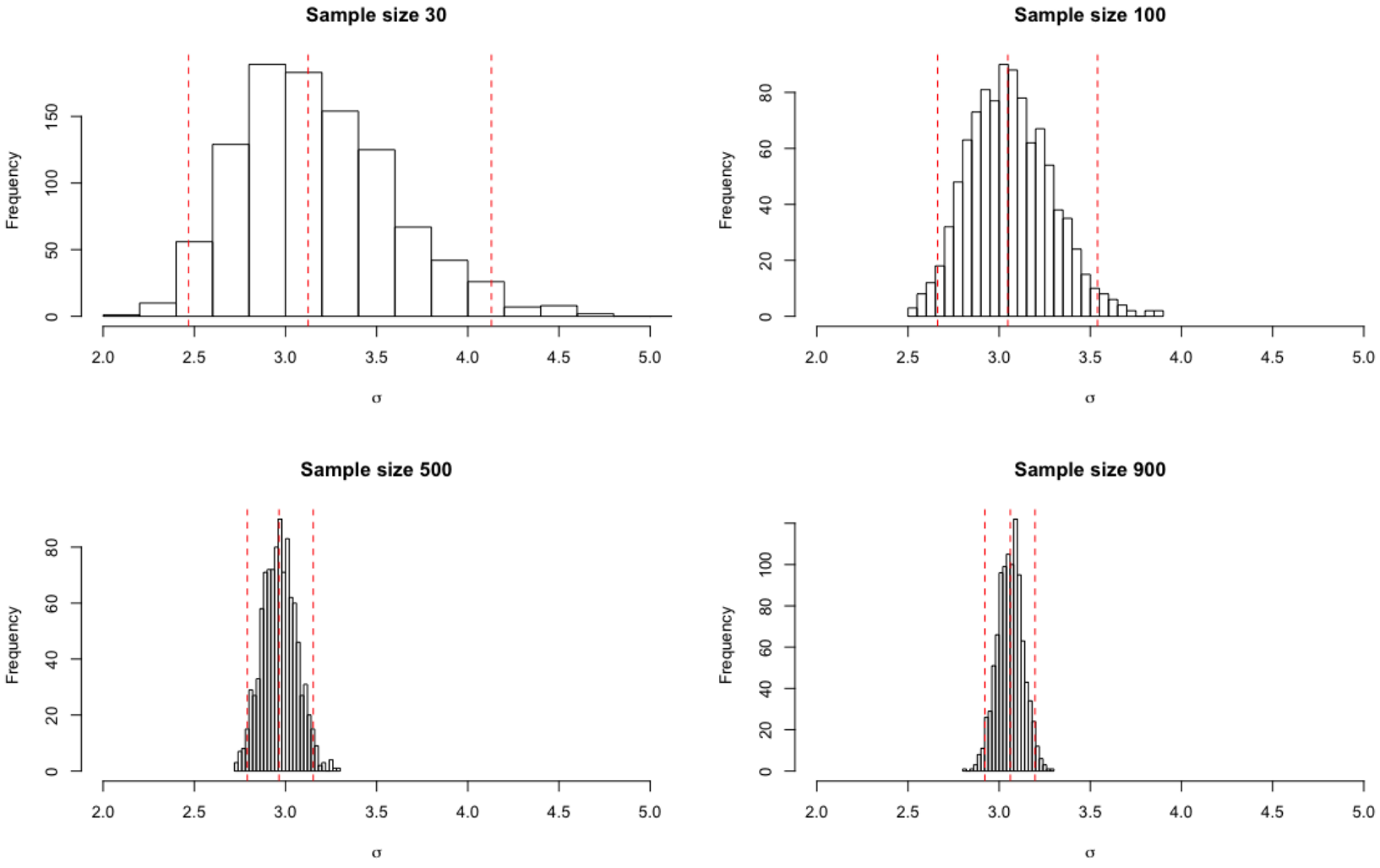
$$\sigma^2|x \sim Scaled\ Inv - \chi^2\left(n, S^2\right)$$

Estimar σ Normal, con μ conocida





Estimar σ Normal, con μ conocida



Modelos uniparamétricos: Poisson

Modelo Poisson

- $x|\theta \sim P(\theta)$: $p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!}$, $x = 0, 1, 2, \ldots$
- $x = (x_1, x_2, \ldots, x_n)$ i.i.d., $p(x|\theta) = \prod_{i=1}^n \frac{1}{x_i!}\theta^{x_i} e^{-\theta} \propto \theta^{t(x)} e^{-n\theta}$
- $t(x) = \sum_{i=1}^n x_i = n\bar{x}$ *estadístico suficiente*
- Prior conjugada $Gamma(\alpha, \beta)$: $p(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$, con hiperparámetros α, β
- Distr. Post. $\theta|x \sim Gamma\,(\alpha + n\bar{x}, \beta + n)$

Exponencial

- $x|\theta \sim Exp(\theta) = Gamma(1, \theta)$: $p(x|\theta) = \theta e^{-x\theta}$, $x > 0$ y $\theta = 1/E(x|\theta)$ *razón*
- Prior conjugada $Gamma(\alpha, \beta)$: $p(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$
- Distr. Post. $\theta|x \sim Gamma\,(\alpha + 1, \beta + x)$: $p(x|\theta) = \theta^n e^{-n\bar{x}\theta}$

Datos epidemiológicos

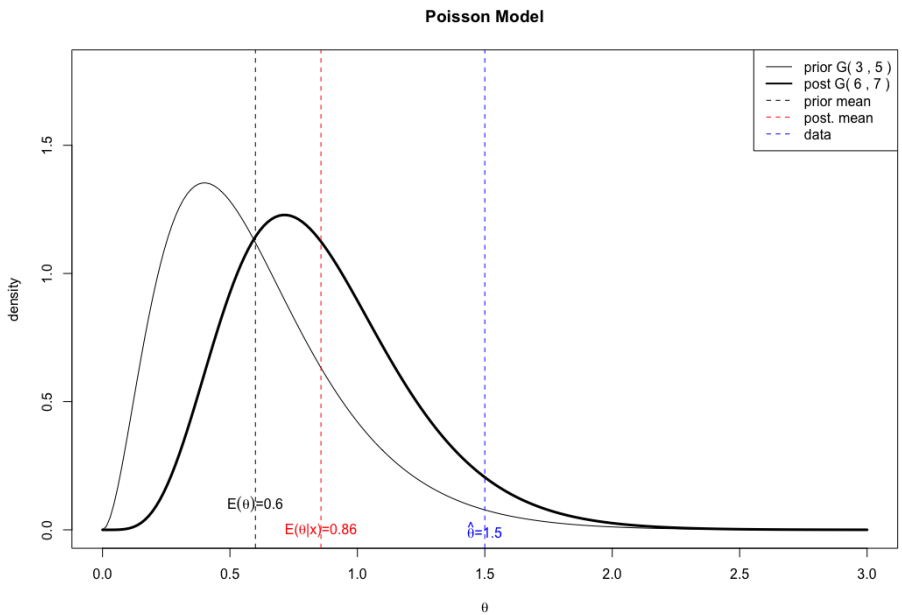
- Estudio del número de muertes por una determinada enfermedad, en una determinada población (ciudad, ...) durante 1 año.
- Datos: 3 muertes por dicha enfermedad en una población de 200.000 hab. $\rightarrow \hat{\theta} = 1.5$ cada 100.000 hab.. al año.
- Modelo Poisson en términos de proporción, θ = tasa de mortalidad por cada 100.000 hab. al año, y exposición, e=200.000 hab:
  - $x \sim P(e\theta)$: $p(x \mid \theta) = \frac{1}{x!}(e\theta)^x exp(-e\theta)$
  - Datos: $x = 3$, con $e = 2.0$. Y $E(\theta) = 0.6$ (estudios)
- Prior conj. $\theta \sim Gamma(\alpha, \beta)$: $p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}exp(-\beta\theta)$
- Posterior $\theta \mid x \sim Gamma(x + \alpha, e + \beta)$:

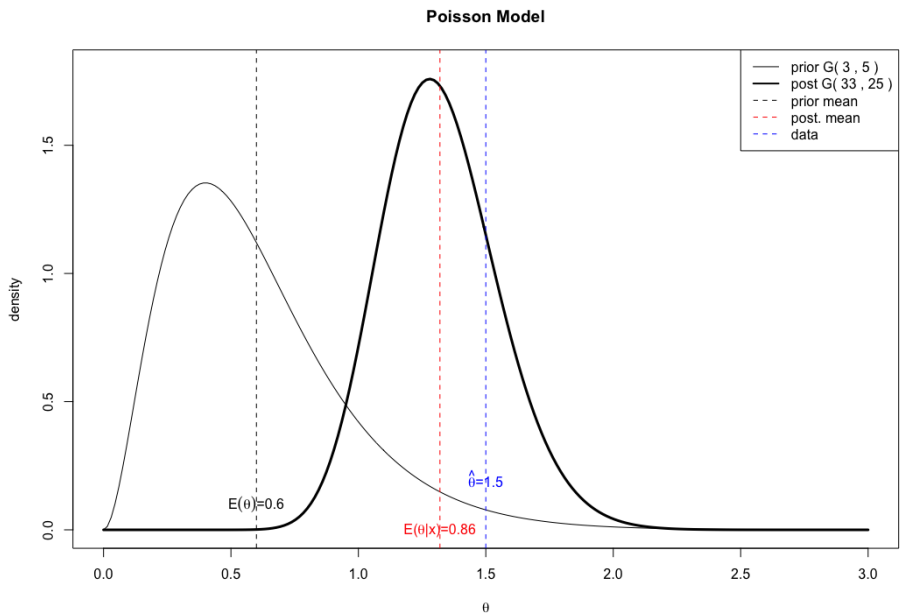
$$p(\theta \mid x) \propto \underbrace{\theta^{\alpha-1}exp(-\beta\theta)}_{\text{prior}}\underbrace{\theta^{x}exp(-e\theta)}_{\text{verosimilitud}} = \theta^{x+\alpha-1}exp(-(e+\beta)\theta)$$

Datos epidemiológicos

- Datos: $x = 3$, con $e = 2.0$;
  $E(\theta) = 0.6 \rightarrow Gamma(3, 5)$:
  Post. $Gamma(6, 7)$,
  $E(\theta|x) \approx 0.86$, $P(\theta > 1 \mid x) \approx 0.3$

- Últimos 10 años: $x = 30$, con $e = 20$.
  Asumiendo cte. pob. y $\theta = {}^{1.5}/_{100.000 hab}$,
  e independencia entre un año y otro:
  Post. $Gamma(33, 25)$,
  $E(\theta|x) \approx 1.32$, $P(\theta > 1 \mid x) \approx 0.93$





- Datos: $\{(x_i, e_i)\}_{i=1,\ldots,n}$, pero θ cte. (indep)
- Prior no inform, $p(\theta) \propto 1/\theta$: $\theta \mid x \sim Gamma(\sum_{i=1}^n x_i, \sum_{i=1}^n e_i)$:

$$p(\theta \mid x) \propto p(\theta) \prod_{i=1}^n p(x_i \mid \theta) \propto \theta^{(\sum_{i=1}^n x_i)-1} exp\left(-\theta \sum_{i=1}^n e_i\right)$$

  $\hookrightarrow$ estimación hiperparams: $\tilde{\alpha} = \sum_{i=1}^n x_i$, $\tilde{\beta} = \sum_{i=1}^n e_i$ (Mod. jerárquicos)
- Prior conj $\theta \sim Gamma(\alpha, \beta)$: $\theta \mid x \sim Gamma(\alpha + \sum_{i=1}^n x_i, \beta + \sum_{i=1}^n e_i)$:

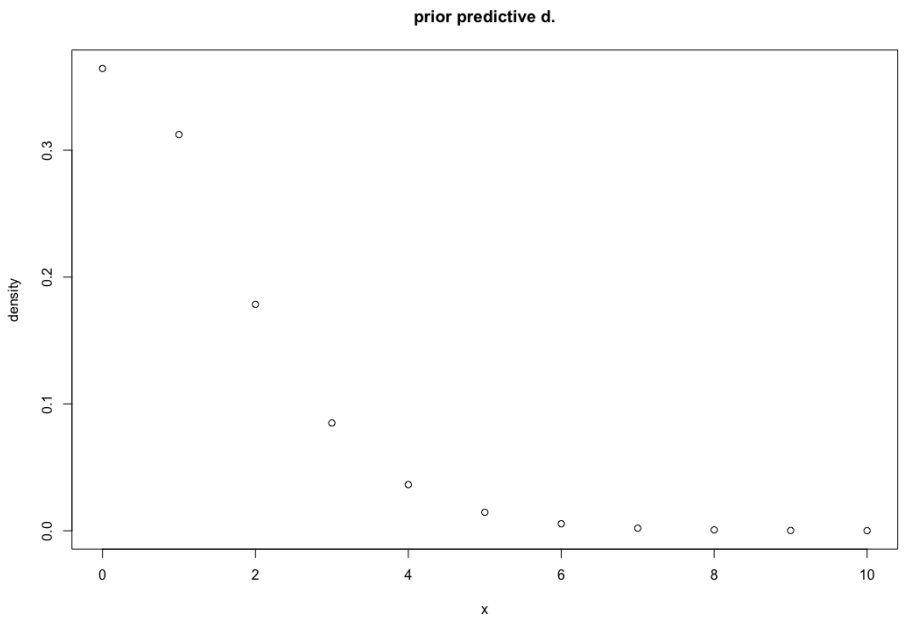
$$p(\theta \mid x) \propto p(\theta) \prod_{i=1}^n p(x_i \mid \theta) \propto \theta^{\alpha-1+\sum_{i=1}^n x_i} exp\left(-(\beta + \sum_{i=1}^n e_i)\theta\right)$$

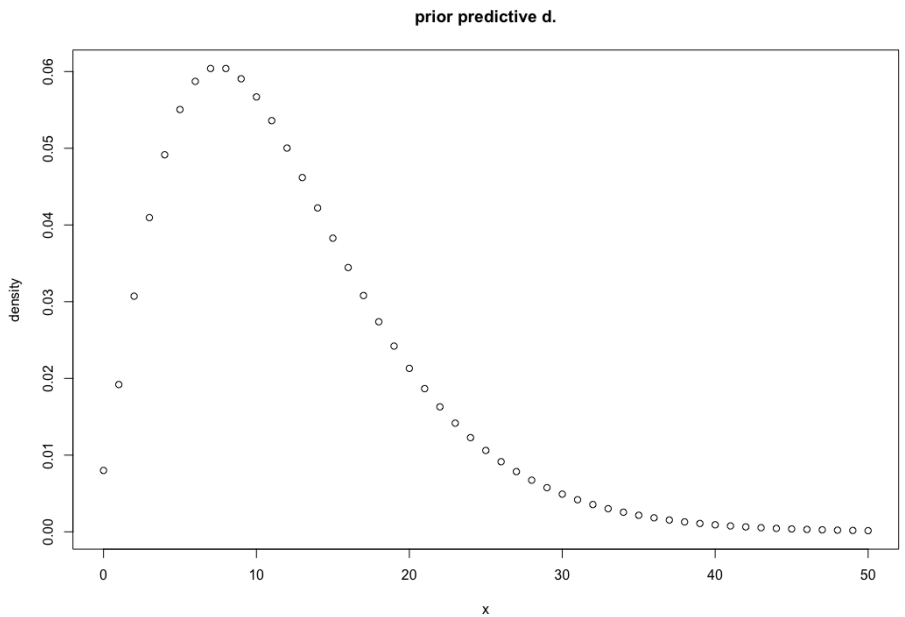
- Validez del modelo $\rightarrow$ D. Prior predictiva
  1. Marginalizando d. conj: $p(x) = \int p(\theta, x) d\theta = \int p(\theta) p(x \mid \theta) d\theta$
  2. Factorizando d. conj: $p(x) = \dfrac{p(\theta) p(x \mid \theta)}{p(\theta \mid x)}$

Datos epidemiológicos

Prior pred. d: $p(x) = \dfrac{Gamma(\theta; 3, 5)Poisson(x \mid e\theta)}{Gamma(\theta; 3 + x, 5 + e)}$

- Datos: $x = 3$, con $e = 2.0$;
- Últimos 10 años: $x = 30$, con $e = 20$.





Modelos multiparamétricos

- Sea (θ_1, θ_2), su fdd conjunta

$$p(\theta_1, \theta_2|x) \propto p(\theta_1, \theta_2)p(x|\theta_1, \theta_2)$$

- Si θ_1 param. de interés, su fdd post. es

$$\begin{aligned} p(\theta_1|x) &= \int p(\theta_1, \theta_2|x)d\theta_2 \\ &= \int p(\theta_1|\theta_2, x)p(\theta_2|x)d\theta_2 \end{aligned}$$

combinación de d.post. cond., dado θ_2, $p(\theta_1|\theta_2, x)$; y la d. marg. post. $p(\theta_2|x)$ (evidencia de los datos y el modelo prior) como f. ponderación/peso

Modelos multiparamétricos: Normal, prior no informativa

- $x \sim N(\mu, \sigma)$, $\{x_1, \ldots, x_n\}$ mas, iid
- Prior no inform., indep. y uniforme en $(\mu, log\sigma) \propto 1$, $\Leftrightarrow$ $p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$ (impropia)
- D. post. conjunta de $\mu, \sigma^2|x$

$$
\begin{aligned}
(a) \quad p(\mu, \sigma^2|x) &\propto p(\mu, \sigma^2) \cdot p(x|\mu, \sigma^2) \\
&\propto (\sigma^2)^{-1} \cdot \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&= \sigma^{-n-2} exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right]\right) \\
&= \sigma^{-n-2} exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s_c^2 + n(\bar{x} - \mu)^2\right]\right)
\end{aligned}
$$

donde $\bar{x}$ y s_c^2 estadísticos suficientes

$$
(b) \quad p(\mu, \sigma^2|x) = p(\mu|\sigma^2, x)p(\sigma^2|x)
$$

Modelos multiparamétricos: Normal, prior no informativa

- D. post. cond de $\mu|\sigma^2, x$

$$
\begin{aligned}
p(\mu|\sigma^2, x) &\propto p(\mu) \cdot p(x|\mu, \sigma^2) \\
&\propto 1 \cdot \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\
&= \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right]\right) \\
&= \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s_c^2 + n(\bar{x} - \mu)^2\right]\right) \\
&\propto exp\left(-\frac{1}{2\sigma^2} n(\mu - \bar{x})^2\right)
\end{aligned}
$$

$$\mu|\sigma^2, x \sim N(\bar{x}, \sigma^2/n)$$

Modelos multiparamétricos: Normal, prior no informativa

- D. post. marg. de $\sigma^2|x$

$$
\begin{aligned}
p(\sigma^2|x) &= \int p(\mu,\sigma^2|x)d\mu \\
&\propto \int \sigma^{-n-2} exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s_c^2 + n(\bar{x}-\mu)^2\right]\right) d\mu \\
&= \sigma^{-n-2} exp\left(-\frac{(n-1)s_c^2}{2\sigma^2}\right) \underbrace{\int exp\left(-\frac{1}{2\sigma^2/n}(\mu-\bar{x})^2\right) d\mu}_{\sqrt{2\pi\sigma^2/n}} \\
&\propto (\sigma^2)^{-(n+1)/2} exp\left(-\frac{(n-1)s_c^2}{2\sigma^2}\right)
\end{aligned}
$$

$$
\boxed{\sigma^2|x \sim Inv-\chi^2(n-1,s_c^2)} \quad \equiv \quad \sigma^2 = \frac{(n-1)s_c^2}{x},\ x \sim \chi^2_{n-1}
$$

NOTA: Análogo estad. suf. $\frac{(n-1)s_c^2}{\sigma^2} \sim \chi^2_{n-1}$

Modelos multiparamétricos: Normal, prior no informativa

- Muestreo de la d. post. conjunta: $p(\mu, \sigma^2|x) = p(\mu|\sigma^2, x)p(\sigma^2|x)$
  1. Muestreamos σ^2 de $\sigma^2|x \sim Inv - \chi^2(n-1, s_c^2)$
  2. Dado σ^2, muestreamos μ de $\mu|\sigma^2, x \sim N(\bar{x}, \sigma^2/n)$
- Es uno de los pocos prob. multiparamétricos lo suficientemente sencillos como para resolver analíticamente.

Modelos multiparamétricos: Normal, prior no informativa

- Forma analítica de la d. post. marginal de μ:

$$
\begin{aligned}
p(\mu|x) &= \int_0^\infty p(\mu,\sigma^2|x)d\sigma^2 \\
&\propto \int_0^\infty \sigma^{-n-2} exp\underbrace{\left(-\frac{1}{2\sigma^2}\overbrace{\left[(n-1)s_c^2+n(\bar{x}-\mu)^2\right]}^{A}\right)}_{z=A/(2\sigma^2)} d\sigma^2 \\
&= \int_0^\infty \left(\frac{A}{2z}\right)^{-\frac{(n+2)}{2}} e^{-z}\frac{A}{2z^2}dz = \left(\frac{A}{2}\right)^{-\frac{n}{2}} \underbrace{\int_0^\infty z^{\frac{n}{2}-1}e^{-z}dz}_{\Gamma(n/2)} \\
&\propto \left((n-1)s_c^2+n(\bar{x}-\mu)^2\right)^{-\frac{n}{2}} \propto \left(1+\frac{n(\mu-\bar{x})^2}{(n-1)s_c^2}\right)^{-\frac{n}{2}}
\end{aligned}
$$

$$
\boxed{\mu|x \sim t_{n-1}(\bar{x}, s^2/n)} \quad \equiv \quad \left.\frac{\mu-\bar{x}}{\sigma/\sqrt{n}}\right|_x \sim t_{n-1}
$$

NOTA: Análogo estad. suf. (pivote) $\left.\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right|_{\mu,\sigma^2} \sim t_{n-1}$

Modelos multiparamétricos: Normal, prior no informativa

- D. post. predictiva,

$$p(\tilde{x}|x) = \int \int p(\tilde{x}|\mu, \sigma^2, x) p(\mu, \sigma^2|x) d\mu d\sigma$$

  - Muestrear d. post. pred.:
    1. Muestreamos μ, σ^2 de su d. post. conj.
    2. Dado (μ, σ^2), muestreamos $\tilde{x}$ de $N(\mu, \sigma^2)$
  - Forma analítica (análogo a $\mu|x$): $\tilde{x}|x \sim t_{n-1}\left(\bar{x}, \left(1 + \frac{1}{n}\right)^{1/2} s\right)$
  - Factorización $p(\tilde{x}|\sigma^2, x) = \int p(\tilde{x}|\mu, \sigma^2, x) p(\mu|\sigma^2, x) d\mu$, conduce a : $p(\tilde{x}|\sigma^2, x) \equiv N\left(\bar{x}, (1 + \frac{1}{n})\sigma^2\right)$ (misma que $\mu|\sigma^2, x$)

Modelos multiparamétricos: Normal, prior no informativa

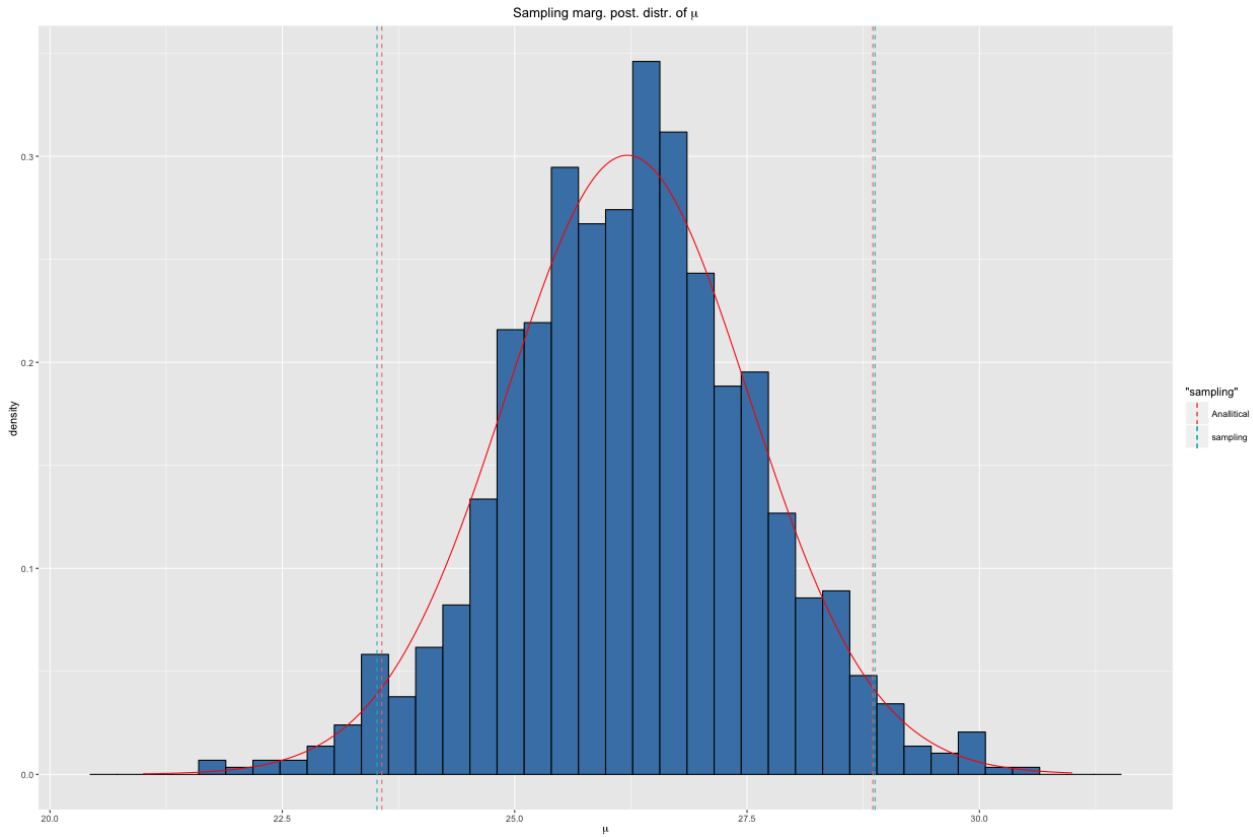
Ejemplo: Estimando la velocidad de la luz

- Simon Newcomb, 1882. Midió el tiempo que tardaba la luz en recorrer una distancia de 7442 m.
- Fichero *'light.txt'* contiene 66 mediciones de Newcomb (desviaciones de 24800 nanoseg). Tenemos dos medidas inusualmente bajas.
- Asumimos (no mejor modelo) distribución normal, mediciones indep.

- $x_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, 66$
- Obj. es inferir $\mu|x$
- Prior no inf.
  $p(\mu, \sigma^2) \propto 1/\sigma^2 \quad \Rightarrow \quad \frac{\mu - \bar{x}}{S/\sqrt{n}}\Big|_x \sim t_{n-1} \quad \bar{x} = 26.2, S = 10.8$
- Analíticamente: $IC_{95\,\%}(\mu) \equiv \bar{x} \pm t_{n-1,0.975}\frac{S}{\sqrt{n}} = [23.6, 28.8]$

Modelos multiparamétricos: Normal, prior no informativa

- Simulación, $p(\mu, \sigma^2|x) = p(\mu|\sigma^2, x)p(\sigma^2|x)$
  1. Muestreamos σ^2 de $\sigma^2|x \sim Inv - \chi^2(n-1, s_c^2)$
     $\Rightarrow \sigma^2 = \frac{65S^2}{x}$, $x \sim \chi^2_{65}$
  2. Dado σ^2, muestreamos μ de $\mu|\sigma^2, x \sim N(\bar{x}, \sigma^2/n)$



Modelos multiparamétricos: Normal, prior conjugada

- $x \sim N(\mu, \sigma)$, $\{x_1, \ldots, x_n\}$ mas, iid. Verosimilitud:

$$\begin{aligned} p(x|\mu,\sigma^2) &\propto \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\ &= \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right]\right) \\ &= \sigma^{-n} exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s_c^2 + n(\bar{x}-\mu)^2\right]\right) \end{aligned}$$

- Prior conjugada: $p(\mu,\sigma^2) = p(\sigma^2)p(\mu|\sigma^2)$, donde $\sigma^2 \sim \chi^2(\nu_0, \sigma_0^2)$ y $\mu|\sigma^2 \sim N(\mu_0, \sigma^2/\kappa_0)$ (marginal μ es t-student)

$$p(\mu,\sigma^2) \propto \sigma^{-1}(\sigma^2)^{-\nu_0/2+1} exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu_0-\mu)^2\right]\right)$$

$$(\mu,\sigma^2) \sim \text{N-Inv-}\chi^2\left(\overbrace{\underbrace{\mu_0}_{local.}, \underbrace{\sigma_0^2/\kappa_0}_{escala}}^{\mu}; \overbrace{\underbrace{\nu_0}_{gl}, \underbrace{\sigma_0^2}_{escala}}^{\sigma}\right)$$

Modelos multiparamétricos: Normal, prior conjugada

- D. post. conjunta

$$
\begin{aligned}
p(\mu,\sigma^2|x) &\propto \sigma^{-1}(\sigma^2)^{-\nu_0/2+1} exp\left(-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2+\kappa_0(\mu_0-\mu)^2\right]\right)\times \\
&\quad \times(\sigma^2)^{-n/2} exp\left(-\frac{1}{2\sigma^2}\left[(n-1)s_c^2+n(\bar{x}-\mu)^2\right]\right) \\
&\quad \dots \\
&\propto \sigma^{-1}(\sigma^2)^{-\nu_n/2+1} exp\left(-\frac{1}{2\sigma^2}\left[\nu_n\sigma_n^2+\kappa_n(\mu_n-\mu)^2\right]\right)
\end{aligned}
$$

i.e. $\boxed{(\mu,\sigma^2|x)\sim \text{N-Inv-}\chi^2\left(\mu_n,\sigma_n^2/\kappa_n;\nu_n,\sigma_n^2\right)}$ donde,

$$
\begin{aligned}
\mu_n &= \tfrac{\kappa_0}{\kappa_0+n}\mu_0+\tfrac{n}{\kappa_0+n}\bar{x} \\
\kappa_n &= \kappa_0+n \\
\nu_n &= \nu_0+n \\
\nu_n\sigma_n^2 &= \nu_0\sigma_0^2+(n-1)s^2+\tfrac{\kappa_0 n}{\kappa_0+n}(\bar{x}+n)^2
\end{aligned}
$$

Modelos multiparamétricos: Normal, prior conjugada

- D. post. condicional $\mu|\sigma^2, x$ es prop. a d. post. conj. $\mu, \sigma^2|x$ con $\sigma^2 = cte$.

$$p(\mu|\sigma^2, x) \quad \propto \quad exp\left(-\frac{1}{2\sigma^2}\kappa_n(\mu_n - \mu)^2\right)$$

i.e. $\boxed{(\mu|\sigma^2, x) \sim N\left(\mu_n, \sigma^2/\kappa_n\right)}$ donde,

$$\left.\begin{array}{rcl} \mu_n & = & \frac{\kappa_0\mu_0}{\kappa_0+n} + \frac{n\bar{x}}{\kappa_0+n} = \frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{\kappa_0}{\sigma^2}\bar{x}}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}} \\ \frac{\sigma^2}{\kappa_n} & = & \frac{1}{\frac{\kappa_0}{\sigma^2}+\frac{n}{\sigma^2}} \end{array}\right\} \equiv \mu|x, \quad \sigma \text{ conocida}$$

- Expresión analítica:

$$\begin{aligned} p(\mu|x) &= \int p(\mu, \sigma^2|x)d\sigma^2 \\ &\propto exp\left(1 + \frac{\kappa_n(\mu_n - \mu)^2}{\nu_n\sigma_n^2}\right)^{-(\nu_n+1)/2} \\ &= t_{\nu_n}\left(\mu_n, \sigma_n^2/\kappa_n\right) \end{aligned}$$

Modelos multiparamétricos: Normal, prior conjugada

- D. post. marg. de $\sigma^2|x$

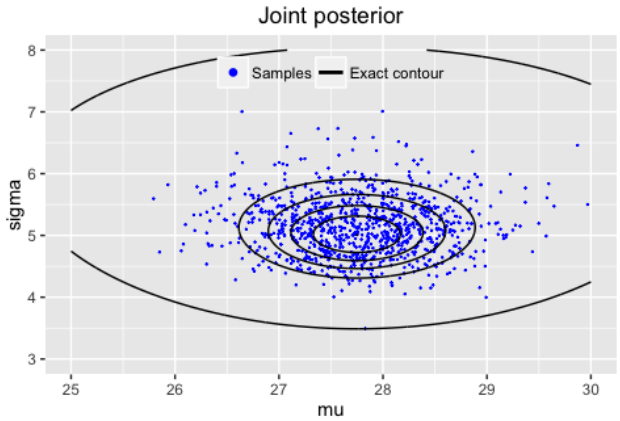
$$
\begin{aligned}
p(\sigma^2|x) &= \int p(\mu,\sigma^2|x)d\mu \\
&\propto \int \sigma^{-1}(\sigma^2)^{-\nu_n/2+1} exp\left(-\frac{1}{2\sigma^2}\left[\nu_n\sigma_n^2+\kappa_n(\mu_n-\mu)^2\right]\right)d\mu \\
&= \sigma^{-1}(\sigma^2)^{-\nu_n/2+1} exp\left(-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right) \underbrace{\int exp\left(-\frac{1}{2\sigma^2}\kappa_n(\mu-\mu_n)^2\right)d\mu}_{\sqrt{2\pi\sigma^2/\kappa_n}} \\
&\propto (\sigma^2)^{-\nu_n/2+1} exp\left(-\frac{\nu_n\sigma_n^2}{2\sigma^2}\right)
\end{aligned}
$$

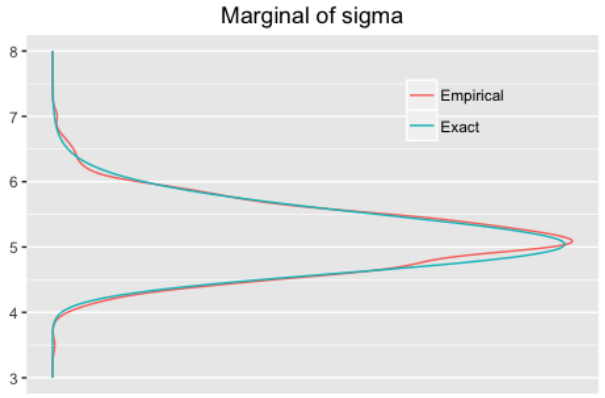
$$
\boxed{\sigma^2|x \sim \text{Inv-}\chi^2(\nu_n,\sigma_n^2)} \quad \equiv \quad \sigma^2 = \frac{\nu_n\sigma_n^2}{x},\ x \sim \chi^2_{\nu_n}
$$

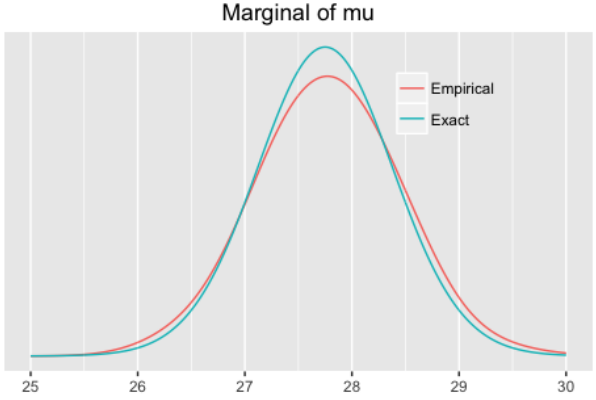
- Muestreo d. post. conj.

$$
p(\mu,\sigma^2|x) = p(\mu|\sigma^2,x)p(\sigma^2|x) \qquad \begin{cases} 1) & \sigma^2|x \sim \text{Inv-}\chi^2(\nu_n,\sigma_n^2) \\ 2) & \mu|\sigma^2,x \sim N\left(\mu_n,\sigma^2/\kappa_n\right) \end{cases}
$$

Ejemplo: Estimando la velocidad de la luz







Modelos multiparamétricos: Modelo Lineal Generalizado

Ejemplo: Experimento biológico

- Desarrollo de medicamentos y componentes químicos, test de toxicidad o biológicos sobre animales o personas. Respuestas suelen ser dicotómicas (muerto/vivo, tumor/no tumor)
- Sea un experimento sobre 20 animales, 5 por cada una de las 4 dosis distintas aplicadas. Los resultados se representan por

$$(x_i, n_i, y_i), \quad i = 1, \ldots, k$$

donde x_i es la i-ésima de $k = 4$ dosis (en escala log.), aplicada sobre n_i animales, con y_i la respuesta obtenida.

Dosis (log g/ml) x_i	No. animales n_i	No. muertes y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

Modelos multiparamétricos: Modelo Lineal Generalizado

- Modelo: 2 params, no conjugado
- Asumimos y_i se distribuyen como una binomial: $y_i|\theta_i \sim B(n_i, \theta_i)$
- θ_i la prob. de muerte para una dosis x_i
- Asumimos independencia de resultados entre grupos, dado $\theta_1, \ldots, \theta_4$. Prior no informativa: $p(\theta_1, \ldots, \theta_4) \propto 1$ (distr. post. θ_i beta)
- Relación dosis-respuesta: modelo *regresión logística* (pues $\theta_i \in [0, 1]$)

$$logit(\theta_i) = \alpha + \beta x_i$$

donde $logit(\theta_i) = log(\frac{\theta_i}{1-\theta_i})$

Distrib. post. conjunta de param. (α, β)

$$\begin{aligned} p(\alpha, \beta | y, n, x) &\propto p(\alpha, \beta | n, x) p(y | \alpha, \beta, n, x) \\ &\propto p(\alpha, \beta | n, x) \prod_{i=1}^{n} p(y_i | \alpha, \beta, n_i, x_i) \end{aligned}$$

Prior

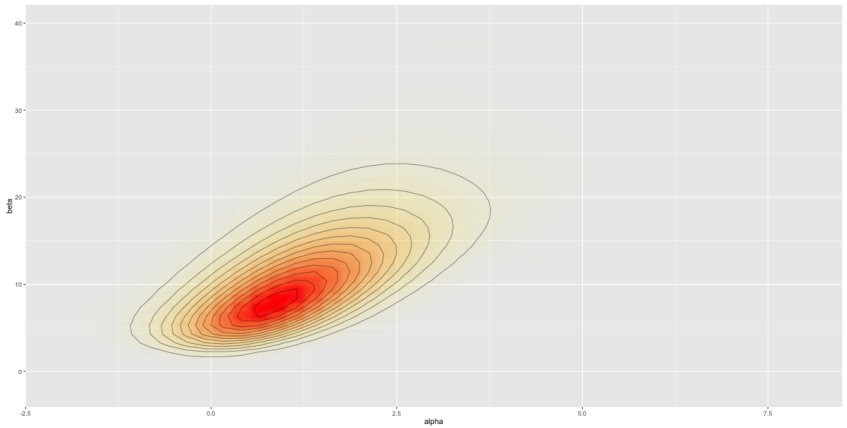
$$p(\alpha, \beta) \propto 1$$

Verosimilitud

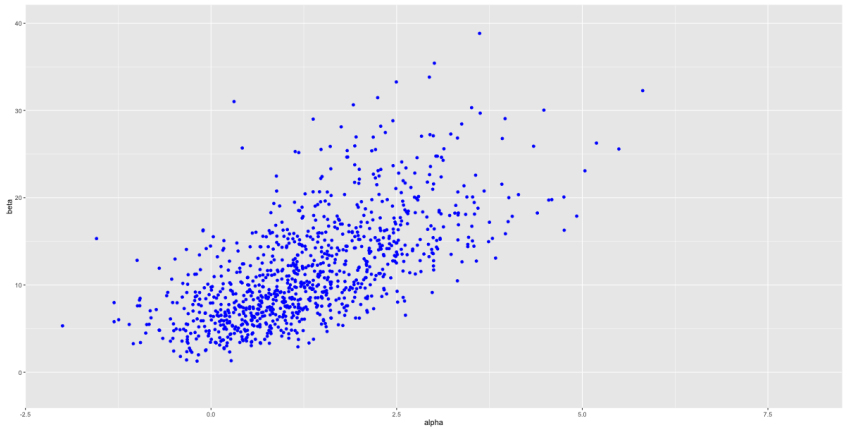
$$p(y_i | \alpha, \beta, n_i, x_i) \propto [logit^{-1}(\alpha + \beta x_i)]^{y_i} [1 - logit^{-1}(\alpha + \beta x_i)]^{n_i - y_i}$$

$$\begin{aligned} \log \mathcal{L} &\propto \sum_{i=1}^{n} y_i \log \left(\frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \right) + (n_i - y_i) \log \left(1 - \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \right) \\ &\propto \sum_{i=1}^{n} y_i (\alpha + \beta x_i) - n_i \log \left(1 + e^{\alpha+\beta x_i} \right) \end{aligned}$$

- Primera estimación param: $(\hat{\alpha}, \hat{\beta}) = (0.8 \pm 1.0, 7.7 \pm 4.9)$
- Iso-contornos densidad post. de (α, β):



- Muestreo de distr. post. conj. de (α, β):



Modelos Jerárquicos

- Mod. multi-paramétricos, relacionados o conectados de cierta forma con la estructura del problema.
- D. prob conj de los params debe reflejar dicha dependencia
- D. prior tal que θ_j's entendidos como una mas de una d. prob común

$$\underbrace{\phi}_{\text{hiperparam}} \rightarrow \underbrace{\theta}_{\text{param}} \rightarrow \underbrace{x}_{\text{obs}}$$

- Mod. multi-paramétricos: pocos params no suf. para ajustar grandes bases de datos, y demasiados params tienden a sobreestimar.
- Mod. jerárquicos proporcionan params suf. para ajustar bien datos (evitando problemas de sobreestimación), además de utilizar distr. pop. que refleje la dependencia entre los params.

Intercambiabilidad

- Hipt. priori, asumir que $\theta = (\theta_1, \ldots, \theta_k)$ es intercambiable: distr. de θ permanece inalterable ante permutaciones de las componentes sus θ_j
- Ninguna otra inform, salvo los datos (sin orden ni agrupamientos)
- θ_j son m.a.s. de d. prior cond. a *hiper-parámetros* ϕ:

$$p(\theta \mid \phi) = \prod_{j=1}^{J} p(\theta_j \mid \phi)$$

  ϕ desconocido $\rightarrow$ tiene su propia distr. de prob.
- Th. Finetti: en el límite $J \rightarrow \infty$

$$p(\theta) = \int \left(\prod_{j=1}^{J} p(\theta_j \mid \phi) \right) p(\phi) d\phi$$

Modelos Jerárquicos Bayesianos

- $x = (x_1, \ldots, x_n)$ intercambiables, en términos de verosimilitud n obs. i.i.d. dados params θ: $p(x \mid \theta)$
- θ_j intercambiables, m.a.s. de d. prior cond. a ϕ: $p(\theta \mid \phi)$
- Clave mod. jerárquico: ϕ desconocido, tiene su propia d. $p(\phi)$
- D. prior conjunta:

$$p(\theta, \phi) = p(\phi)p(\theta \mid \phi)$$

- D. post conjunta:

$$p(\theta, \phi \mid x) \propto p(\theta, \phi)p(x \mid \theta, \phi) = p(\phi)p(\theta \mid \phi)p(x \mid \theta)$$

Modelos Jerárquicos conjugados

- D. post. cond de θ dado ϕ, para datos obs. x fijos:

$$p(\theta \mid \phi, x) = \prod_{j=1}^{J} p(\theta_j \mid \phi, x)$$

  analíticamente o sim. MCMC

- D. post. marginal de ϕ:

$$\begin{aligned} p(\phi \mid x) &= \int p(\phi, \theta \mid x) d\theta \quad \text{marginalizar d. conj} \\ &= \frac{p(\phi, \theta \mid x)}{p(\theta \mid \phi, x)} \quad \text{algebraicamente} \end{aligned}$$

Modelo riesgo de tumor en un grupo de ratas de laboratorio

- Experimentos Tarone 1982 ($I = 71$): $x_i = $ n° ratas con tumor, $n_i = $ n° total de ratas, y θ_i prob. tener tumor, $i = 1, \ldots, I$.
- Hipts: $x_i \sim B(n_i, \theta_i)$ indp. $i = 1, \ldots, I$
- $\theta_i \sim Beta(\alpha, \beta)$ m.a.s.
- (α, β) hiperparámetros desconocidos:
  - Método momentos:

$$E(\theta) = \frac{\alpha}{\alpha+\beta} \quad Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{E(\theta)(1-E(\theta))}{\alpha+\beta+1}$$

  - Mod jerárquico: $p(\alpha, \beta)$ hiper-prior no informativa
- D. post. conj

$$p(\theta, \alpha, \beta \mid x) \propto \overbrace{p(\alpha,\beta)}^{\text{hiperprior}} \overbrace{p(\theta \mid \alpha, \beta)}^{Beta(\alpha,\beta)} \overbrace{p(x \mid \theta, \alpha, \beta)}^{Bin(n,\theta)}$$

$$\propto p(\alpha,\beta) \prod_{i=1}^{I} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \prod_{i=1}^{I} \theta_i^{x_i}(1-\theta_i)^{n_i-x_i}$$

- D. post. cond de θ dado α, β, para datos obs. x fijos:

$$\begin{aligned} p(\theta \mid \alpha, \beta, x) &= \prod_{i=1}^{I} p(\theta \mid \alpha, \beta, x) = \\ &= \prod_{i=1}^{I} \frac{\Gamma(\alpha+\beta+n_i)}{\Gamma(\alpha+x_i)\Gamma(\beta+n_i-x_i)} \theta_i^{\alpha+x_i-1}(1-\theta_i)^{\beta+n_i-x_i-1} \end{aligned}$$

- D. post. marginal de α, β:

$$\begin{aligned} p(\alpha, \beta \mid x) &= \frac{p(\theta, \alpha, \beta \mid x)}{p(\theta \mid \alpha, \beta, x)} \\ &\propto p(\alpha, \beta) \prod_{i=1}^{I} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x_i)\Gamma(\beta+n_i-x_i)}{\Gamma(\alpha+\beta+n_i)} \end{aligned}$$

- Elección de la hiperprior ?:
  - Reparametrizamos, $\theta \sim Beta(\alpha, \beta)$ con $\alpha, \beta > 0$:

    $E(\theta) = \frac{\alpha}{\alpha+\beta} \in (0,1) \quad \Rightarrow \quad \phi_1 = logit\left(\frac{\alpha}{\alpha+\beta}\right) = log\left(\frac{\alpha}{\beta}\right) \in \mathcal{R}$

    $`n` = \alpha + \beta > 0 \quad \Rightarrow \quad \phi_2 = log\,(\alpha+\beta) \in \mathcal{R}$

    Asumimos indep: $p(\phi_1, \phi_2) = p(\phi_1)p(\phi_2)$
  - D. uniforme $p\left(log\left(\frac{\alpha}{\beta}\right), log(\alpha+\beta)\right) \propto 1 \rightarrow$ d. post. impropia (!!) :

$$
\begin{aligned}
p(x \mid \alpha, \beta) &\propto \prod_{i=1}^{I} \frac{[\alpha \ldots (\alpha + x_i - 1)][\beta \ldots (\beta + n_i - x_i - 1)]}{(\alpha+\beta)\ldots(\alpha+\beta+n_i-1)} \\
&= \prod_{i=1}^{I} \frac{[\frac{\alpha}{\alpha+\beta} \ldots (\frac{\alpha}{\alpha+\beta} + \frac{x_i-1}{\alpha+\beta})][\frac{\beta}{\alpha+\beta} \ldots (\frac{\beta}{\alpha+\beta} + \frac{n_i-x_i-1}{\alpha+\beta})]}{(1+\frac{1}{\alpha+\beta})\ldots(1+\frac{n_i-1}{\alpha+\beta})} \\
(\alpha+\beta) \rightarrow \infty,\ \frac{\alpha}{\beta}\ \textbf{cte} \quad &\approx \prod_{i=1}^{I} \left(\frac{\alpha}{\alpha+\beta}\right)^{x_i} \left(\frac{\beta}{\alpha+\beta}\right)^{n_i-x_i} \approx cte
\end{aligned}
$$

d. prior determina si d. post tiene integral finita en este límite

Problema gral. en Mod. Jerárquicos: prior uniforme para log desv. std. de params, resulta d. post. impropia.
Forma de solventarlo: prior unif. sobre la desv. std. del param, no su log

- Prior unif. (indpt) en $p\left(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}\right) \propto 1$:
  - $p(\alpha,\beta) \propto (\alpha+\beta)^{-5/2}$, impropia
  - $p\left(log\left(\frac{\alpha}{\beta}\right), log(\alpha+\beta)\right) \propto \alpha\beta(\alpha+\beta)^{-5/2}$, no inform, impropia pero dominada por verosimilitud $\Rightarrow$ d. post. propia

$$(\phi_1, \phi_2) = g(\alpha, \beta) \rightarrow \begin{array}{l} p_\phi(\phi_1, \phi_2) = p_{\alpha,\beta}(g^{-1}(\phi_1, \phi_2)) \left| \frac{d}{d\phi} g^{-1}(\phi_1, \phi_2) \right| \\ p(\alpha, \beta) = p_\phi(g(\alpha, \beta)) \left| \frac{d}{d(\alpha,\beta)} g(\alpha, \beta) \right| \end{array}$$

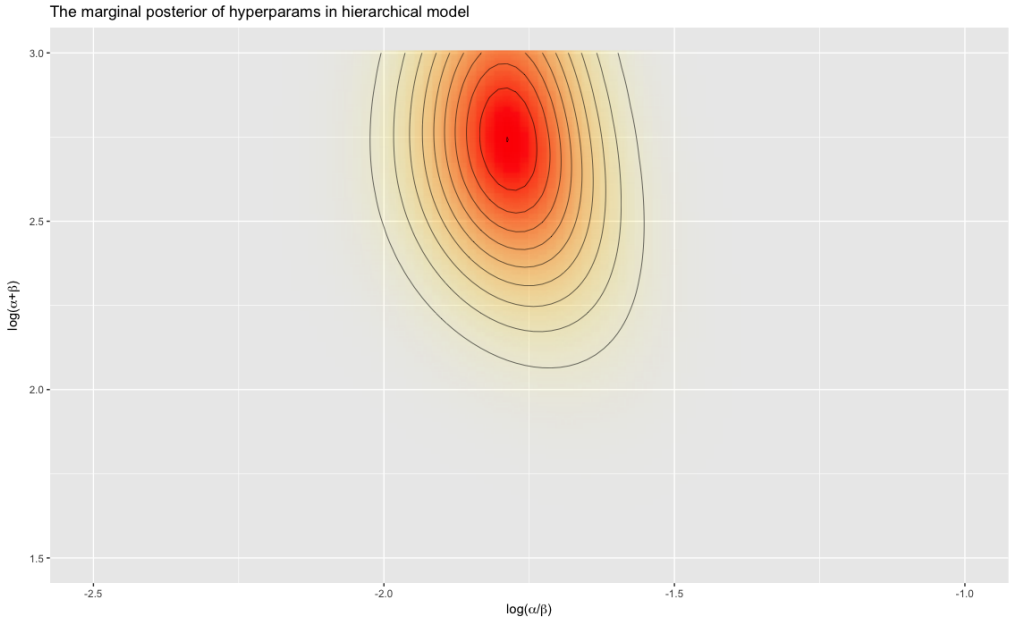
- D. post. marginal de hiperparam $(\overbrace{log\,(\alpha/\beta)}^{\phi_1}, \overbrace{log(\alpha+\beta)}^{\phi_2})$:

$$p\left(log\left(\frac{\alpha}{\beta}\right), log(\alpha+\beta) \mid x\right) \propto \frac{\alpha\beta}{(\alpha+\beta)^{5/2}} \prod_{i=1}^{I} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+x_i)\Gamma(\beta+n_i-x_i)}{\Gamma(\alpha+\beta+n_i)}$$

- Aprox. inicial (MM):

$$\begin{cases} E(\theta) \simeq \bar{x} = 0.136 \\ Var(\theta) \simeq S^2 = 0.103 \end{cases}$$

$$\begin{cases} (\alpha_0, \beta_0) = (1.4, 8.6) \\ \\ (log(\frac{\alpha_0}{\beta_0}), log(\alpha_0+\beta_0)) = \\ \qquad = (-1.8, 2.3) \pm 2dex \end{cases}$$



- Cálculo de la D. post. marginal:
  1. *Grid*: $\{\phi_{1m}, \phi_{2n}\}$ que cubra la d. post.
  2. *Aprox. discreta de la d. post sobre grid* (log, precisión numérica):

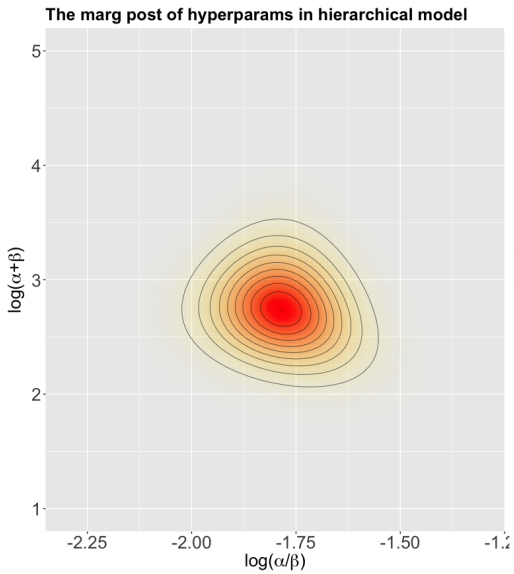
$$p(\phi_{1j}, \phi_{2k} \mid x) \simeq \frac{p\left(log\left(\frac{\alpha_j}{\beta_k}\right), log(\alpha_j + \beta_k)\right) p(x \mid \alpha_j, \beta_k)}{\sum_{m,n} p\left(log\left(\frac{\alpha_m}{\beta_n}\right), log(\alpha_m + \beta_n)\right) p(x \mid \alpha_m, \beta_n)}$$

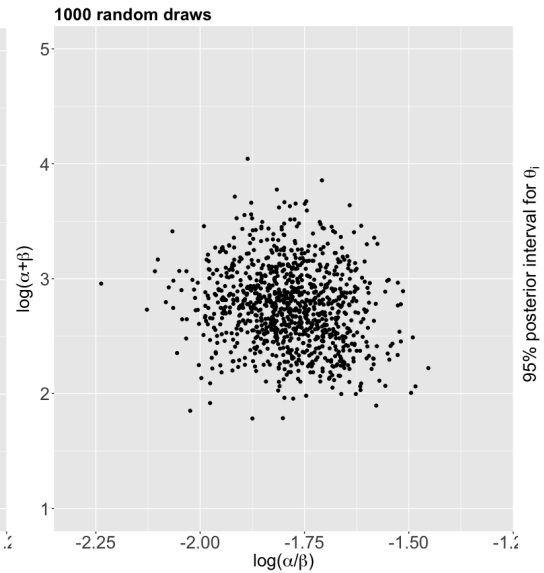
$$= \frac{\frac{\alpha_j \beta_k}{(\alpha_j+\beta_k)^{5/2}} \prod_{i=1}^{I} \frac{\Gamma(\alpha_j+\beta_k)}{\Gamma(\alpha_j)\Gamma(\beta_k)} \frac{\Gamma(\alpha_j+x_i)\Gamma(\beta_k+n_i-x_i)}{\Gamma(\alpha_j+\beta_k+n_i)}}{\sum_{m,n} \frac{\alpha_m \beta_n}{(\alpha_m+\beta_n)^{5/2}} \prod_{i=1}^{I} \frac{\Gamma(\alpha_m+\beta_n)}{\Gamma(\alpha_m)\Gamma(\beta_n)} \frac{\Gamma(\alpha_m+x_i)\Gamma(\beta_n+n_i-x_i)}{\Gamma(\alpha_m+\beta_n+n_i)}}$$

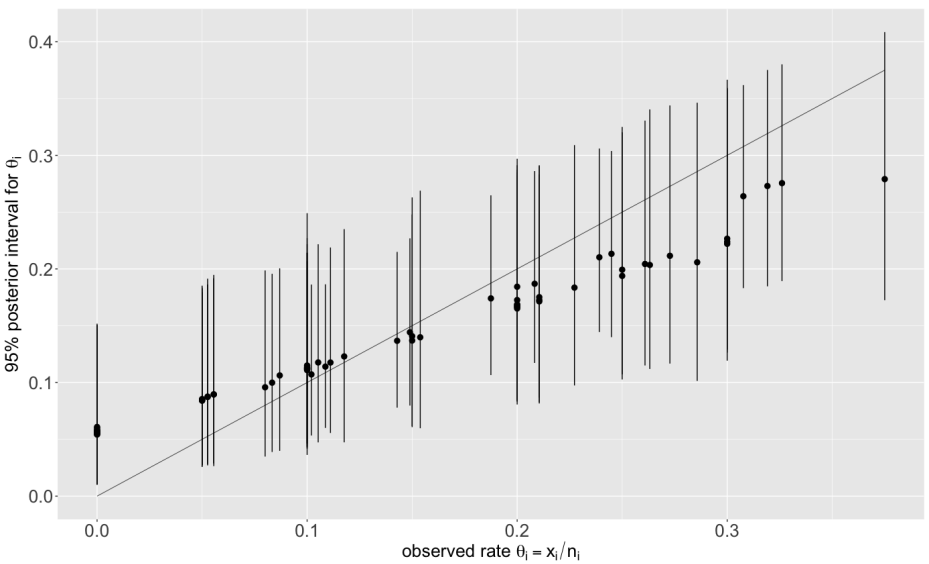
$$E(\alpha \mid x) \simeq \sum_{m,n} \alpha_m p\left(log\left(\frac{\alpha_m}{\beta_n}\right), log(\alpha_m + \beta_n) \mid x\right) = 2.4$$

$$E(\beta \mid x) \simeq \sum_{m,n} \beta_n p\left(log\left(\frac{\alpha_m}{\beta_n}\right), log(\alpha_m + \beta_n) \mid x\right) = 14.3$$

- Cálculo de la D. post. cond:
  3. Tomamos m.a.s. $\{(\phi_1^k, \phi_2^k) \mid x\}_{k=1,\ldots,1000} \Rightarrow \{(\alpha^k, \beta^k) \mid x\}_{k=1,\ldots,1000}$
  4. $\forall j = 1, \ldots, J$: $\theta_j \mid \alpha, \beta, x \sim Beta(\alpha + x_j, \beta + n_j - x_j)$







Computación Bayesiana

- Objetivo por excelencia del análisis Bayesiano es la distribución posterior de los parámetros, $p(\theta|x) \propto p(\theta)p(x|\theta)$

$$log\ p(\theta|x) \propto log\ p(\theta) + \sum_{i=1}^{n} log\ p(x_i|\theta)$$

- donde $\theta = (\theta_1, \ldots, \theta_k)$ params. desconocidos, con prior $p(\theta)$
- información $x = (x_1, \ldots, x_n)$ mas, iid $p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$
- En general los problemas principales son:
  1. Cómo **generar muestras aleatorias de la d. post**
  2. Cómo **calcular integrales con respecto a la d. post**

Computación Bayesiana

- Los principales métodos que vamos a estudiar son:
  1. Integración de Montecarlo
  2. Rejection
  3. Muestreo Importante
  4. Markov Chain Monte Carlo Method (MCMC)
  5. Muestreo de Gibbs
  6. Metropolis Hasting

Modelo beta-binomial

- Estudio mortalidad por cáncer de estómago en 20 ciudades de Missouri, $x = $ n° muertes por cáncer, $n = $ n° indiv. en riesgo
- Paquete R `LearnBayes`: `cancermortality` (Datos muy dispersos)
- Hipts: $x \sim Beta - bin(n, \alpha, \beta)$, i.e. $x \sim B(n, \theta)$ donde $\theta \sim Beta(\alpha, \beta)$

$$\begin{aligned} p(x) &= \int_0^1 p(x|\theta)p(\theta)d\theta \\ &= \frac{\binom{n}{x}}{B(\alpha,\beta)} \underbrace{\int_0^1 \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}d\theta}_{B(x+\alpha, n-x+\beta)} \\ &= \binom{n}{x} \frac{B(\kappa\eta + x, \kappa(1-\eta) + n - x)}{B(\kappa\eta, \kappa(1-\eta))} \end{aligned}$$

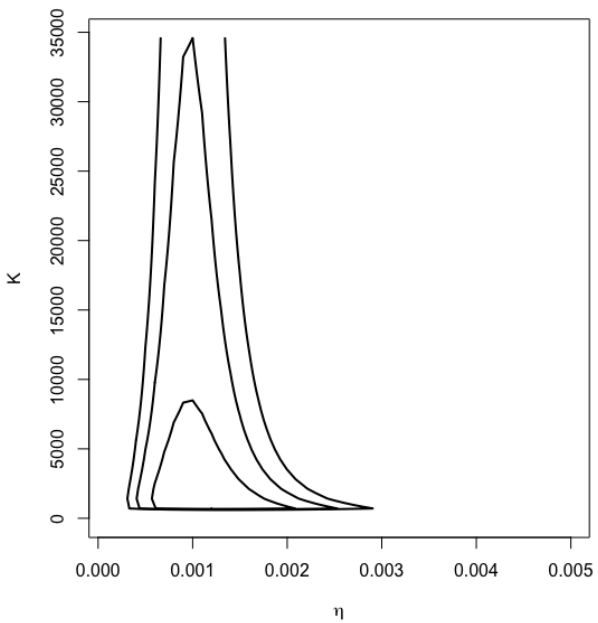
donde $0 < \eta = \frac{\alpha}{\alpha+\beta} = E(\theta) < 1$, y $\kappa = \alpha + \beta > 0$ (tamaño muestral priori)

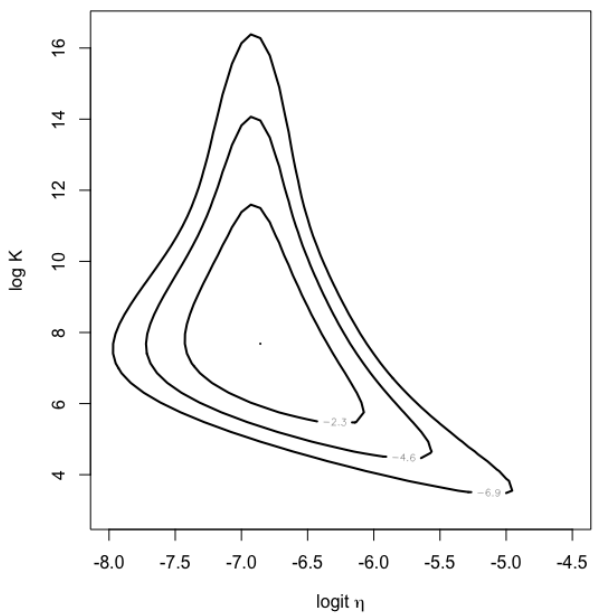
$E(x) = n\frac{\alpha}{\alpha+\beta} = n\eta \qquad Var(x) = \frac{n\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)} = n\eta(1-\eta)\frac{\kappa+n}{\kappa+1}$

Modelo beta-binomial

- D. priori (no inform): $p(\eta, \kappa) \propto \frac{1}{\eta(1-\eta)} \frac{1}{(1+\kappa)^2}$
- D. post $p(\eta, \kappa|x) \propto \frac{1}{\eta(1-\eta)} \frac{1}{(1+\kappa)^2} \prod_{i=1}^{20} \frac{B(\kappa\eta+x_i, \kappa(1-\eta)+n_i-x_i)}{B(\kappa\eta, \kappa(1-\eta))}$
- Re-param: $(\theta_1, \theta_2) = (logit(\eta), log(\kappa))$

$$p(\theta_1, \theta_2|x) \quad \propto \quad p_{\eta,\kappa}\left(\frac{e^{\theta_1}}{1+e^{\theta_1}}, e^{\theta_2}\middle| x\right) \frac{e^{\theta_1+\theta_2}}{(1+e^{\theta_1})^2}$$





Aproximación basada en la Moda Posterior

- Método resumir d. post. $p(\theta|x)$ basado en comportamiento de la densidad sobre su moda
- Sea $h(\theta) = log\,(p(\theta)p(x|\theta))$, y $\hat{\theta} = Mo(\theta|x)$. Aprox. de Taylor en $\hat{\theta}$,

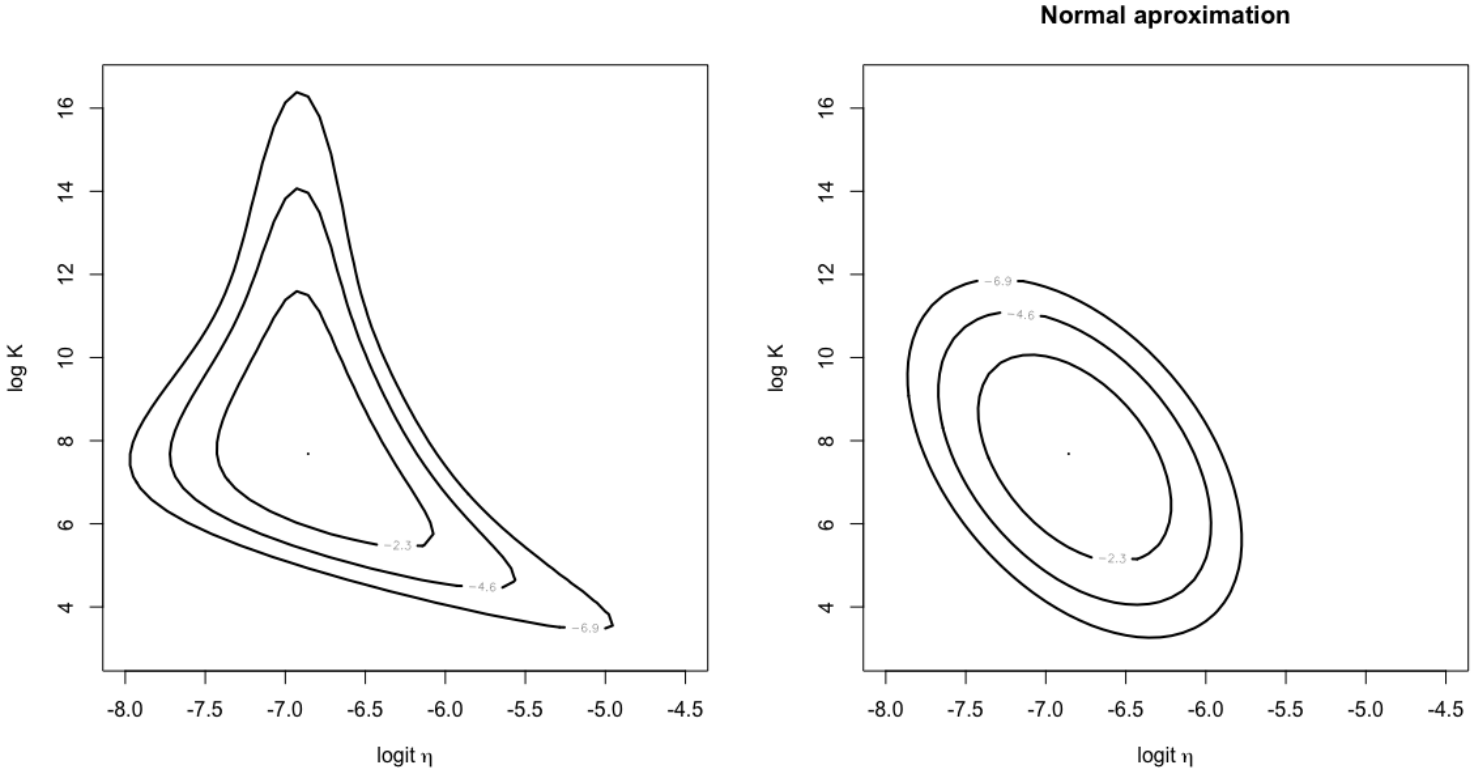
$$h(\theta) \approx h(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T h''(\hat{\theta})(\theta - \hat{\theta})$$

  i.e. $\theta|x \sim N(\hat{\theta}, -h''(\hat{\theta})^{-1})$
- Determinación moda $\hat{\theta}$: M. Newton, Alg. Nelder-Mead (`laplace`, θ^0)
- D. prior predictiva

$$p(x) \approx \int h(\theta, x)d\theta = \int h(\theta)d\theta \approx (2\pi)^{d/2} p(\hat{\theta})p(x|\hat{\theta})|-h''(\hat{\theta})|^{1/2}$$

Modelo beta-binomial

- $\theta^0 = (-7, 6) \rightarrow \hat{\theta} = (-6.82, 7.58), \Sigma = \begin{pmatrix} 0.1579 & -0.2970 \\ -0.2970 & 2.6966 \end{pmatrix}$
- $IP_{90\,\%}(logit\eta) = (-7.282, -6.358),\ IP_{90\,\%}(log\kappa) = (5.666, 9.486)$



Integración de Montecarlo

- Método gral. de resumir d. post. basado en simulaciones
- Sea $\theta^1, \ldots, \theta^m$ una muestra aleatoria de $p(\theta|x)$, y $h(\theta)$ func. cualquiera de los params.
- Media post de $h(\theta)$ se aproxima por la media muestral,

$$E(\theta|x) = \int h(\theta)p(\theta|x)d\theta \simeq \frac{1}{m}\sum_{k=1}^{m} h(\theta^k) = \bar{h}$$

con error std asociado

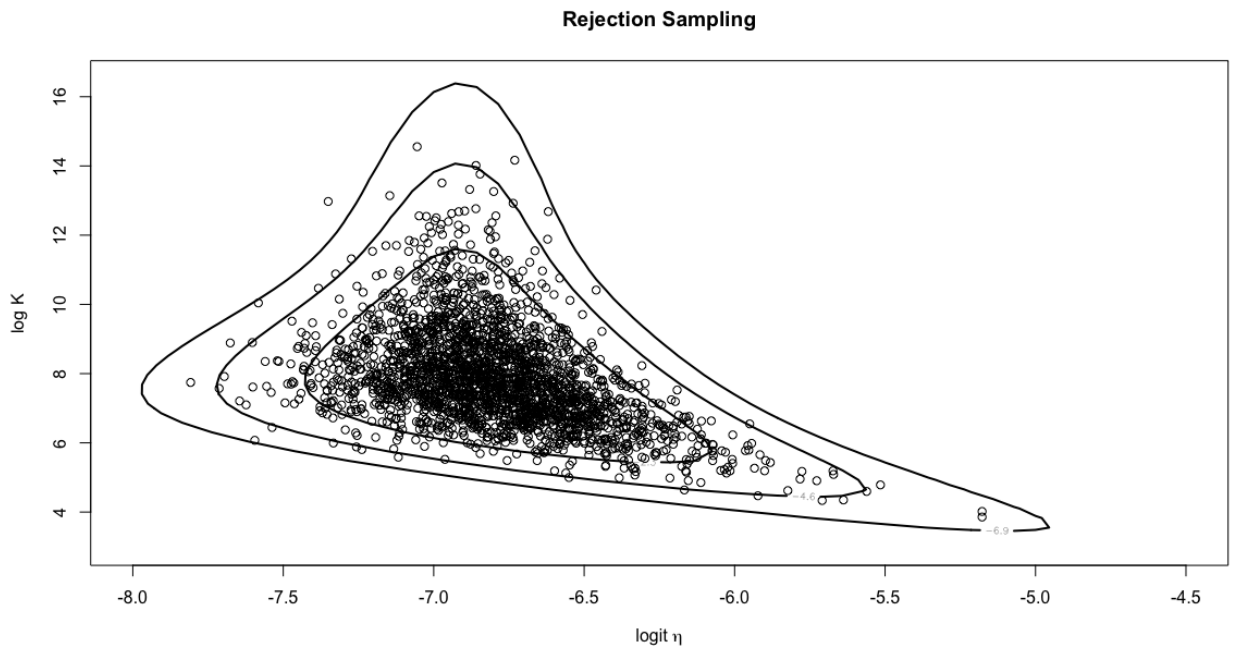
$$se_{\bar{h}} = \sqrt{\frac{\sum_{k=1}^{m}(h(\theta^k) - \bar{h})^2}{(m-1)m}} = \frac{s}{\sqrt{m}}$$

Muestreo de Rechazo

- Queremos generar una muestra de d. post. $\mathbf{p}(\theta|\mathbf{x})$, con forma func. arbitraria, e incluso podemos desconocer la cte de normalización
- Supongamos que podemos encontrar una f de densidad $\tilde{p}(\theta)$ tal que:
  - Sea fácil obtener realizaciones de $\tilde{p}$
  - Sea similar a la d. post en términos de media y dispersión
  - Existe una contante c tal que: $p(\theta|x) \leq c\tilde{p}(\theta)$
- Idea intuitiva: Si f es una distr. con soporte en $[a, b]$, y $f(x) \leq m, \forall x$. El objetivo es simular la distribución f:
  - Sean $X \sim U[a, b]$ e $Y \sim U[0, m]$, si $y < f(x)$ acepte a x como una simulación de $f(x)$
  - I.e, las simulaciones uniformes en $[a, b] \times [0, m]$ de $X \times Y$ las acepta únicamente cuando está bajo el área de la densidad f
- Algoritmo general:
  1. Simule de forma independiente u de distr. $U[0, 1]$ y θ de la distr. $\tilde{p}$.
  2. Si $u \leq \frac{p(\theta|x)}{c\tilde{p}(\theta)}$ acepte la simulación. Caso contrario se rechaza
  3. Repetir el procedimiento hasta obtener tamaño muestral deseado

Modelo beta-binomial

- $\theta = (\theta_1, \theta_2) = (logit(\eta), log(\kappa))$
- D. aprox: $\tilde{p} \equiv t_{\nu=4}\left(\hat{\theta} = (-6.82, 7.58), \Sigma = \begin{pmatrix} 0.1579 & -0.2970 \\ -0.2970 & 2.6966 \end{pmatrix}\right)$
- Cálculo cte : $p(\theta|x) \leq c\tilde{p}(\theta), \forall\theta \Leftrightarrow log(c) \approx \max_\theta logp(\theta|x) - log\tilde{p}(\theta)$



Muestreo Importante

- Normalmente la cte. normalización de d. post $p(\theta|x)$ es desconocida, por lo que dada $h(\theta)$ arbitraria,

$$E(\theta|x) = \frac{\int h(\theta)p(\theta)p(x|\theta)d\theta}{\int p(\theta)p(x|\theta)d\theta} = \frac{\int h(\theta)\omega(\theta)\tilde{p}(\theta)d\theta}{\int \omega(\theta)\tilde{p}(\theta)d\theta}$$

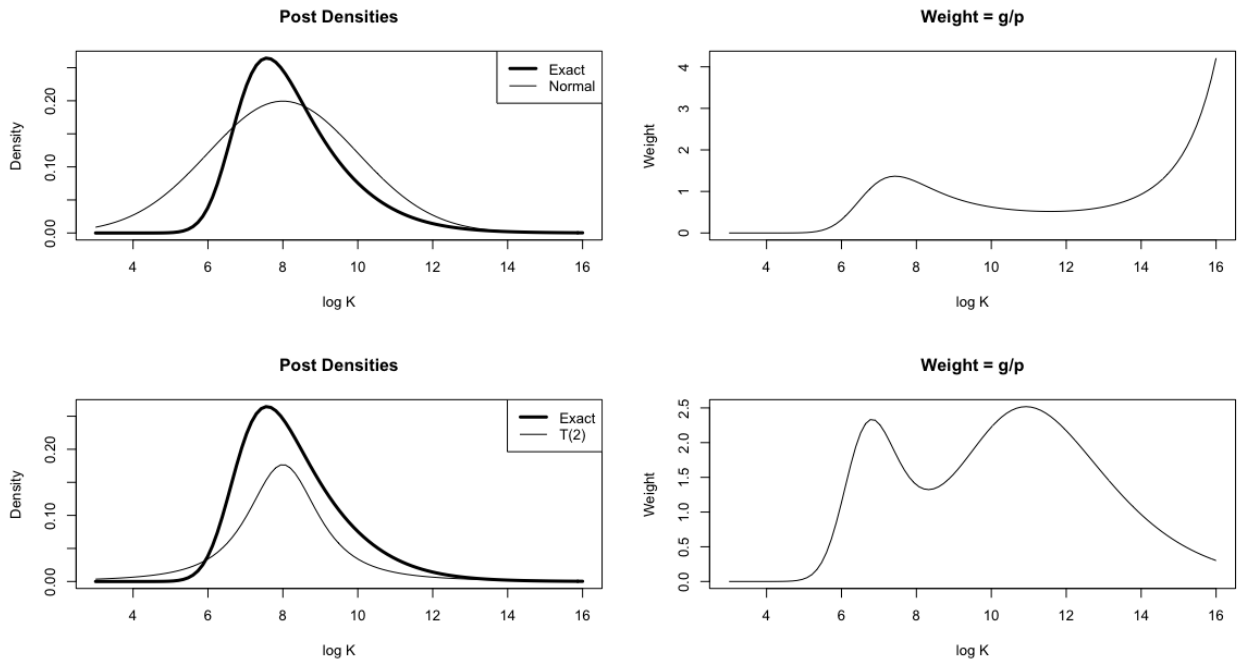
- Si podemos muestrear $\{\theta^k\}_{k=1,\ldots,m}$ directamente de d. post, aprox. dicha esperanza con M. Montercarlo: $E(\theta|x) \sim \bar{h} = \frac{1}{m}\sum_{k=1}^{m} h(\theta^k)$
- En caso contrario, sea $\tilde{p}(\theta)$ una distr. que aproxime d. post, y de la cual es fácil obtener realizaciones: $\{\theta^k\}$.
  - Definimos los pesos, $\omega(\theta^k) = \frac{p(\theta^k)p(x|\theta^k)}{\tilde{p}(\theta^k)}$
  - El alg. de muestreo importante está basado en el siguiente teorema:

$$\bar{h} = \frac{\sum_{k=1}^{m} h(\theta^k)\omega(\theta^k)}{\sum_{k=1}^{m} \omega(\theta^k)} \rightarrow E(\theta|x)$$

  - Error std. asoc. $se_{\bar{h}} = \frac{\sqrt{\sum_{k=1}^{m}(h(\theta^k)-\bar{h})^2\omega(\theta^k)}}{\sum_{k=1}^{m}\omega(\theta^k)}$

Modelo beta-binomial

- $\theta = (\theta_1, \theta_2) = (logit(\eta), log(\kappa))$, ¿$E(\theta_2 \mid x, \theta_1 = \tilde{\theta}_1)$?
- D post: $p(\theta_2|x, \theta_1) \propto \frac{\kappa}{(1+\kappa)^2} \prod_{i=1}^{20} \frac{B(\kappa\eta+x_i, \kappa(1-\eta)+n_i-x_i)}{B(\kappa\eta, \kappa(1-\eta))}$, donde $\eta = \frac{e^{\theta_1}}{1+e^{\theta_1}}$ y $\kappa = e^{\theta_2}$
- $\bar{h} = 7.9623$, $se_{\bar{h}} = 0.0193$



Markov Chain Montecarlo Method

- Un proceso estocástico $\{X_t\}_{t=0,1,\ldots}$, con valores en un conjunto finito $\mathcal{S}$, es un conjunto finito si la secuencia de variables aleatorias satisface la **Propiedad de Markov**:

$$p(X_{t+1} = s|X_0, X_1, \ldots, X_t) = p(X_{t+1} = s|X_t) \quad \forall s \in \mathcal{S}, t \geq 1$$

- Definimos p_0 como la función de distribución (discreta) de X_0 y $P_{i,j} = p(X_{t+1} = s_j|X_t = s_i)$ *matriz de transición*

- Si es posible ir desde cualquier estado a cualquier estado en uno o más pasos decimos que la cadena de Markov es *irreducible*.

- Supongamos que estamos en un estado particular, si solo podemos volver a dicho estado en intervalos regulares decimos que estamos ante una cadena de Markov *periódica*. Aquellas cadenas de Markov que no sean periódicas se denominan *aperiódicas*.

- Sea p_n la distribución no condicional de estar en el periodo n en cada uno de los estados. Entonces: $p_n = p_0 P$

Markov Chain Montecarlo Method

- Una distribución p es *estacionaria* si $p_{t+1} = p_t$ (i.e. $p = pP$)
- Una cadena de Markov es asintóticamente estacionaria si para toda d. inicial p_n, existe $lim_{t\to\infty} p_t = p$, y p es una distribución estacionaria de la cadena.
- Es fácil simular una cadena de Markov si se puede simular la distr. inicial y de la probabilidad de transición (trivial en el caso de cadenas, pero no en el caso general de procesos de Markov).
- La relevancia para el análisis Bayesiano consiste en poder identificar la distribución de interés (posterior) como la distribución estacionaria de una cadena de Markov asintóticamente estacionaria.
- El método de muestreo de Gibbs y Metropolis - Hasting son ejemplos de cadenas de Markov.

Muestreo Gibss

- Dado $\theta = (\theta_1, \ldots, \theta_p)$, en ocasiones no es fácil generar una muestra independiente de d. post. conj. $p(\theta|x)$ para dim. alta
- El siguiente método explota una forma especial de expresar la d. post en función d. cond.,

$$p(\theta|x) = p(\theta_1|\theta_2, x)p(\theta_2|x) = p(\theta_2|\theta_1, x)p(\theta_1|x), \quad \theta = (\theta_1, \theta_2)$$

cuando es más fácil generar una muestra de la d. post. cond.

Algoritmo Gibss

- Suponga que sabemos generar muestras aleatorias de todas las distribuciones posterior condicionales de $p(\theta|x)$
- Seleccione un parámetro de inicio, $\theta^0 = (\theta_1^0, \ldots, \theta_p^0)$, para $s = 1, \ldots, N$:
  - Genere θ_1^s usando $p(\theta_1|x, \theta_{-1}^{s-1})$
  - $p(\theta_2|x, \theta_1^s, \theta_{-\{1,2\}}^{s-1}) \rightarrow \theta_2^s$

    $\vdots$
  - $p(\theta_p|x, \theta_1^s, \theta_{-\{1,2\}}^{s-1}) \rightarrow \theta_p^s$
  - Elimine las primeras S_0 simulaciones para independizar el resultado de la elección inicial de los parámetros.
- De la misma forma que el método de integración de Montecarlo:

$$\bar{g}_S = \frac{1}{S - S_0} \sum_{k=S_0+1}^{S} g(\theta^k) \rightarrow E[g(\theta|x)]$$

Metropolis-Hastings

- Al igual que el método de Gibbs, este algoritmo es un ejemplo de Markov Chain Montecarlo Methods (MCMM), muy útil para muestrear las d. post. Bayesianas.
- Obj: simular una cadena de Markov cuya distr. equilibrio sea d. post.
- En este, la distribución que se utiliza para generar una realización depende de la simulación anterior.
- Tiene también similitudes con el algoritmo de muestreo importante. En este se reduce el peso de simulaciones que generan valores de la densidad, muy diferentes a los que se obtienen de la función de muestreo importante.
- En MH a todas las simulaciones se les da el mismo peso, pero no todas son aceptables.

Algoritmo Metropolis-Hastings

- Sea $\tilde{p}(\theta)$ función de densidad candidata (*proposal* o *jumping* d.), una distr. que aproxime d. post y de la cual es fácil obtener realizaciones: $\{\theta^k\}$ (similar alg. muestreo importante).
- Seleccione un parámetro de inicio, θ^0 $(p(\theta^0|x) > 0)$,
- Para $t = 1, 2, \ldots$
  - Genere θ^* usando $\tilde{p}(\theta^*|\theta^{t-1})$ (*transition kernel*)
  - Calcule la prob. aceptación:

$$\alpha(\theta^{t-1}|\theta^*) = min\left(\frac{p(\theta^*|x)\tilde{p}(\theta^{t-1}|\theta^*)}{p(\theta^{t-1}|x)\tilde{p}(\theta^*|\theta^{t-1})}, 1\right)$$

  - Defina $\theta^t = \begin{cases} \theta^* & \text{con prob. } \alpha(\theta^{t-1}|\theta^*) \\ \theta^{t-1} & \text{cc.} \end{cases}$
  - Elimine las primeras S_0 simulaciones para independizar el resultado de la elección inicial de los parámetros.

Casos particulares: Alg. Metropolis

- Caso particular del Algoritmo Metropolis-Hastings, donde el kernel es **simétrica**: $\tilde{p}(\theta^*|\theta^{t-1}) = \tilde{p}(\theta^{t-1}|\theta^*)$
- En este caso la función de aceptación no depende del kernel,

$$\alpha(\theta^{t-1}|\theta^*) = min\left(\frac{p(\theta^*|x)}{p(\theta^{t-1}|x)}, 1\right)$$

- Ejemplos de densidades candidatas simétricas son: $\tilde{p}(\theta|\theta^{t-1}) = f(|\theta - \theta^{t-1}|)$, donde f cualquier densidad.

Casos particulares: MH independiente, caminata aleatoria

- El caso independiente corresponde al caso en que $\tilde{p}(\theta^{t-1}|\theta) = f(\theta)$
- Esta versión del algoritmo es útil cuando existe una aproximación adecuada de la posterior que podemos usar como kernel (candidato a densidad generadora)
- En este caso el algoritmo se reduce a un algoritmo de muestreo importante
- El caso de la caminata aleatoria corresponde al caso en que $\theta^* = \theta^{t-1} + z$ donde z es una variable aleatoria con cualquier distribución independiente de las variables θ. Si la distribución es simétrica alrededor de cero entonces es un caso de distribución simétrica.
- Este caso es útil cuando no se conoce una aproximación adecuada de la densidad de la posterior.